Improving the Spatial Accuracy of Mobile Positioning Data Based on Fine-Scale Human Mobility Pattern Analysis

By

Li Xu

B.S., Peking University, 2011

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

Of the requirement for the degree of

Master of Arts

Department of Geography

2013

This thesis entitled:
Improving the Spatial Accuracy of Mobile Positioning Data Based on
Fine-scale Human Mobility Pattern Analysis
written by Li Xu
has been approved for the Department of Geography

_____
Professor Kenneth Foote

_____
Assistant Professor Stefan Leyk

_____
Assistant Professor Elisabeth Root

Date_____

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline.

IRB protocol # _____

Xu, Li (MA, Geography, Department of Geography)

Improving the Spatial Accuracy of Mobile Positioning Data Based on Fine-scale Human Mobility Pattern Analysis

Thesis directed by Professor Kenneth E. Foote

Mobile phone, an invention in the 1970s, has become the most widely adopted Information and Communication Technology (ICT) over its forty years history. The advantages of mobile positioning data or call detailed records (CDRs) over Global Positioning Systems (GPS) and traditional techniques such as census or travel surveys make it a popular source in the GIScience, human mobility, and many other studies. However, limitations are also raised by researchers while using mobile positioning data. An important one of them is the relatively low spatio-temporal accuracy. This work thus proposed an approach to fix this limitation, that is, I employed land use categories and Point of Interests (POIs) to improve the spatial accuracy of mobile positionings based on human mobility pattern analysis.

Contents

List of Tables

List of Figures

Chapter 1 Introduction

Since Martin Cooper dials the first mobile call to his competitor at Bell Laboratories in April 1973 at New York, mobile phone technology has become the most ubiquitously and rapidly adopted Information and Communication Technology (ICT) on the planet. Nowadays the developing cellular network has almost taken the place of the "spatial-fixed and temporal-inaccessible" wire telephone (Janelle 1995; Mokhtarian and Meenakshisundaram 1999). People are increasingly relying on mobile phones to have a spatially and temporally flexible lifestyle, especially after the smart phones appeared. As a results, the mobile positioning data or call detailed records (CDRs) offer means of observing both individual and collective level movement, shedding light on revealing human activity patterns. Since human mobility and ICTs have risen rapidly during the past decade, many approaches have been developed based on CDRs to characterize mobility patterns with different focuses (Azevedo et al. 2009; Candia et al. 2008; Gonzalez, Hidalgo, and Barabási 2008; Song et al. 2010). This data have also been widely introduced to numerous fields including social networks (Kwan 2007; Onnela et al., 2007; Eagle, Pentland, and Lazer 2007; Clauset and Eagle 2007), tourism (Ahas et al. 2007a; Ahas et al. 2008), urban planning and transportation (Ratti et al. 2006; Wang et al., 2012), and public health (Wesolowski et al., 2012).

The mobile positioning data are welcome by lots of researchers due to its advantages over traditional data collecting techniques and global positioning systems (GPS). While compared to data collected from traditional methods such as census and travel survey, the mobile positioning data involves spatio-temporal dynamics. Although GPS data have even higher spatio-temporal accuracy, it is very costly and usually works only for small groups of people. The mobile

positioning data, on the other hand, passively records all mobile subscribers and thus involves large volumes of individual data. However, researchers also raised some limitations while utilizing CDRs, including privacy issue and surveillance fear (Minch 2004), data mixing problem, and relatively low spatio-temporal accuracies (Schulz, Bothe, and Körner 2012). This work focuses on fixing one of the limitations, that is, I raised an approach to improve the spatial accuracy of mobile positioning data based on fine-scale human mobility analysis. The methodology was developed with the idea coming from dasymetric mapping.

Chapter 2 Research Context

2.1 The value of mobile phone positioning data

As mentioned in Chapter 1, mobile positioning data has higher spatio-temporal accuracy than traditional techniques do. It also can involve large volumes of individuals simultaneously. These advantages over traditional data collecting methods and GPS have made mobile positioning data a popular source in the GIScience and human mobility studies (Lu and Liu, 2012). For example, despite the fundamental importance of studying human mobility, our understanding of the basic laws governing human travel has been remained limited for a long time. Recently however, significant progress has been made in understanding human travel patterns with the help of mobile phone positioning data. Large-scale, long-term studies on mobile positioning records show a high degree of temporal and spatial regularity in human trajectories, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations (Gonzalez, Hidalgo, and Barabási 2008).

Significant efforts have been made on mobility studies for a great variety of other geography-related topics using mobile positioning data, especially in the field of urban studies. Ahas and Mark (2005), for example, were able to detect the dynamic characteristics of the urban-scale travel of citizens with mobile positioning data for 300 users over seven successive, serving as a supplementation for static urban planning. In their recent article, Silm and Ahas (2010) studied the seasonal variability of population in Estonian municipalities, and developed a methodology for monitoring short-term population mobility. The use of mobile phone data showed great advantages in discovering and monitoring patterns of the population's short-term

mobility over large territories while compared to traditional methods. Ratti et al. (2006) investigated the cell phone and other hand-held devices records at the macro scale in Milan, Italy in 2004. In their work, the authors reported that statistical characteristics of the call intensity revealed important traffic nodes and provided clues to understanding land use patterns.

Another interesting area is associated with health and medical geography issues. Wang et al. (2009) explored the fundamental spread patterns of mobile phone viruses and found that a blue tooth virus can reach all susceptible handsets with enough time, but it spreads slowly because of the constraint by human mobility, offering ample time for developing countermeasures. Since the spatially spreading pattern of blue tooth viruses is proven to be similar to influenza, severe acute respiratory syndrome (SARS), and other contact-based human diseases, it offers insights on the control of outbreak of these human diseases. Emergency management is also a promising application that made use of mobile phone data. Candia et al. reported in their work (2008) that anomalous events such as natural catastrophes and terrorist attacks in human society give rise to spatially extended patterns of mobile phone data that can be meaningfully quantified. These spatio-temporal anomalous fluctuations can be discovered and are relevant in the real-time detection of emergency situations.

2.2 Difficulties of using mobile phone positioning data

Although the use of mobile phone data in studies of human mobility is very successful, it also has limitations such as spatial accuracy problems and privacy issues. In the reviewed studies, problems with spatial accuracy were either ignored, or regarded as having slight impact at a large research scale (Gonzalez, Hidalgo, and Barabási 2008; Candia et al. 2008). Different causes produce two types of spatial uncertainties. The first is the spatial uncertainty associated

with the cellular network and the positioning techniques. Today's various technologies give rise to several mobile positioning approaches and result in different geographical accuracies, which may differ from ten meters to tens of kilometers (Ahonen and Eskelinen 2003). Technical solutions for tracing mobile phone location are also different for different network standards and hardware producers (Adams, Ashwell, and Baxter 2003). However, the most commonly used method is Cell-ID, which uses the unique IDs of individual cell phone base towers. It locates all call events made within a service area to the corresponding base tower, making it impossible to know the precise location of a user within the service cell. The accuracy of Cell-ID thus varies significantly depending on the size of network cells, which ranges from 0.002 to 100 square kilometers (Ahas et al. 2007a).

A second type of spatial uncertainty is produced by the variability in call frequency through time. Since mobile positioning only records data when people use cell phones (sometimes including SMS; Internet or GPRS services; and location-based services), no information is available between telecommunications. The trajectory extracted from mobile positioning data will thus only partially represent actual human movement.

Many efforts have been made on measuring and improving the first type of spatial uncertainty. Ahas et al. (2007b) explored the estimated mobile positioning error in a total of 180,000 measurements of 27,000 mobile phones made over nine months in 2003 in Estonia. The results showed the spatial error of 89 percent of positionings studied remained below 1500 meters in urban areas while it is greater than 2000 meters for 55 percent of the positionings in rural areas. In 2004, they further conducted an experiment with records from 117 mobile phone users. A total of 717 actual locations of mobile phones were measured using GPS, which are chronologically synchronized with the mobile phone positioning results. 52 percent of the

measurements conducted in urban areas had accuracy better than 400 meters and 50 percent of the positionings carried out in rural areas were better than 2600 meters. The results of their work suggest that spatial accuracy of mobile phone data is adequate for commuting and regional studies, but will encounter difficulties to link position to finer scale such as blocks, streets or buildings.

Since findings are limited due to the insufficient spatial accuracy of mobile positioning data, many approaches are proposed by researchers to achieve the most accurate data possible (Ahas et al. 2007a). First, the accuracy will increase if the technical network equipment is improved. Positioning requirements can be taken into account when setting up and directing antennas. A well-known example is A-GPS (Assistant-GPS) that incorporates GPS positioning to mobile phone network and the resulting accuracy can be down to ten meters. However, the wide improvement of infrastructure is a long-term goal that is not suitable for current researches. A second way is to use more advanced mobile positioning technology in the existing infrastructure (Adams, Ashwell, and Baxter 2003). Better methods of positioning have existed for some time already but their deployment is slow as there are not many users of location-based services, and the more accurate methods require greater investments to technological equipment in mobile networks. Finally, positioning can be made more accurate by introducing supplementary questionnaires and interactive travel diaries to mobile phone users. This presents great potential for geographical surveys by providing additional information on the respondent's daily movement habits and routes, but this method is limited to small scale study and privacy issues must be paid more attention to.

2.3 Using ideas of dasymetric mapping

Among the two types of spatial uncertainties introduced in Section 2.2, my work focuses on the first spatial uncertainty that is caused by the Cell-ID. This spatial uncertainty varies significantly with the size of mobile phone service network cells, which can become considerably high especially where the density of base towers is low. While existing methods are limited to improve the spatial accuracy of mobile phone records through better data collection methods, my work addresses the issue by proposing new ways to work with existing data. Although it is impossible to obtain exact positions and reconstruct a completed trajectory through mobile phone data, I developed a method for improving the spatial accuracy to a certain level with the idea coming from dasymetric mapping.

The early development of dasymetric mapping was driven by population mapping studies. A dasymetric map depicts quantitative areal data using boundaries that divide the mapped area into zones of relative homogeneity with the purpose of best portraying the underlying statistical surface (Eicher and Brewer 2001). Instead of using simple areal weighting in traditional areal interpolation, the use of ancillary data to refine mapping results makes dasymetric mapping somehow 'intelligent'. The ancillary data act as a limiting variable or a related variable in the dasymetric mapping process. Limiting variables restrict the possible occurrences of mapped variable in the original unit. That is, a limiting variable sets a maximum percentage of the mapped variable that can occur. Related variables are associated with the variable of interest in self-defined rules. It can be expressed in complex ways but does not necessarily limit the mapped variable. For example, Eicher and Brewer used a 70-20-10 breakdown rule to dasymetric map two population variables and four house value variable in county level: 70 percent of county data were assigned to urban grid cells, 20 percent to

agricultural/woodland cells, and 10 percent to forested cells. The mapping results returned a better estimation of distributions of the six socio-economic variables while validating with block group level data.

In traditional dasymetric mapping, limiting variables and related variables deal with the amount or percent of mapped variable(s) that could occur. In my work, the method was adapted from the traditional ways: the mapped variable became the probability of a telecommunication occurring at a certain area. A supervised land use classification map and a clustering of POIs (Point of Interests) were developed and employed separately as ancillary data, in an attempt to reduce the spatial uncertainty of mobile phone positioning data. Relationships between these two ancillary data and human activity patterns at known positions were established, and were then applied as related variables in a dasymetric mapping-like process to refine the distribution of mobile phone calls within service cells (hereinafter referred to as the *Land-use method* and the *POI method*, respectively).

Chapter 3 Methodology

3.1 Study area

Three primary datasets collected from Beijing were employed in this research, including a GPS survey dataset, a Landsat ETM+ remote sensed imagery, and a POI dataset, which will be introduced in greater details in the following sections. Although the imagery and the POIs data covered the entire Beijing area, most of the human movements extracted from the GPS dataset located at a smaller region, the urban Beijing. The study area for this research was thus narrowed to encompass only the central Beijing area, which is mainly urban. Moreover, The Landsat ETM+ imagery has a scene size of 170km x 183km, which is much larger than the Beijing city. In order to save time and improve the efficiency of raster calculation involved processes, the Landsat scene was subset to the extent of the study area. Two methods were attempted to delimit the study area. A straightforward way is to use the convex hull of all GPS points. It produced a compact study area, thus provided a high calculation efficiency because a convex hull is the convex polygon with the smallest area that contains all involved points. However, this method resulted in an irregular shaped and not very handy working map, especially when GPS outliers or floating points exist.

Therefore, the other method, using the MBR (Minimum Bounding Rectangle) of the central Beijing area, was employed. Figure 1 is an administrative map of Beijing showing the core city districts (red), surrounding urban districts (blue), interior suburban districts (green) and the far north districts and counties (dark yellow). The central Beijing area is defined as the six inner districts: Xicheng District, Dongcheng District, Shijingshan District, Haidian District, Chaoyang District, and Fengtai District. The study area is thus the MBR of the six districts as

shown in Figure 1, which covers a rectangular area of approximately 50.5km by 43.5km.

Figure 1. Study area in Beijing (adapted from Wikimedia Commons)

Before getting into discussing the data and working steps, a workflow chart (Figure 2) could schematically show how different steps are connected, and provides a better overview of the different options I have and steps I did in context.

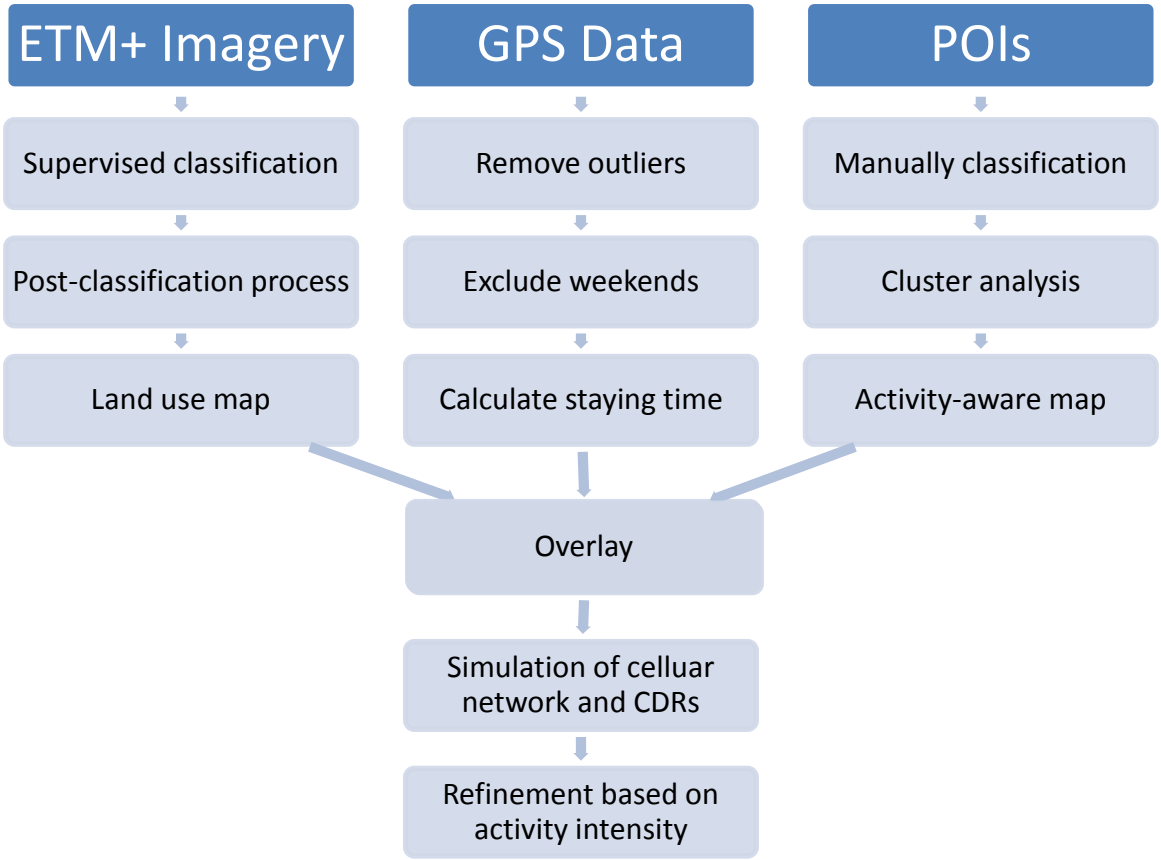

Figure 2. Flowchart of working steps

3.2 GPS Data

The first dataset[1], derived from a survey in Beijing in July 2010 that includes GPS tracking points of one hundred individuals' paths, was used to extract the distribution and intensity of human activities. The one hundred participants were selected from the residents of two areas in Beijing: Tiantongyuan in the north urban area and Yizhuang in the south (see Figure

[1] Data from Professor Yanwei Chai, Department of Urban and Economic Geography, Peking University

3). Each of the survey areas had 50 participants and the time span was nine days for both of them (one is from 2010/07/09 to 2010/07/17; the other is from 2010/07/18 to 2010/07/26).

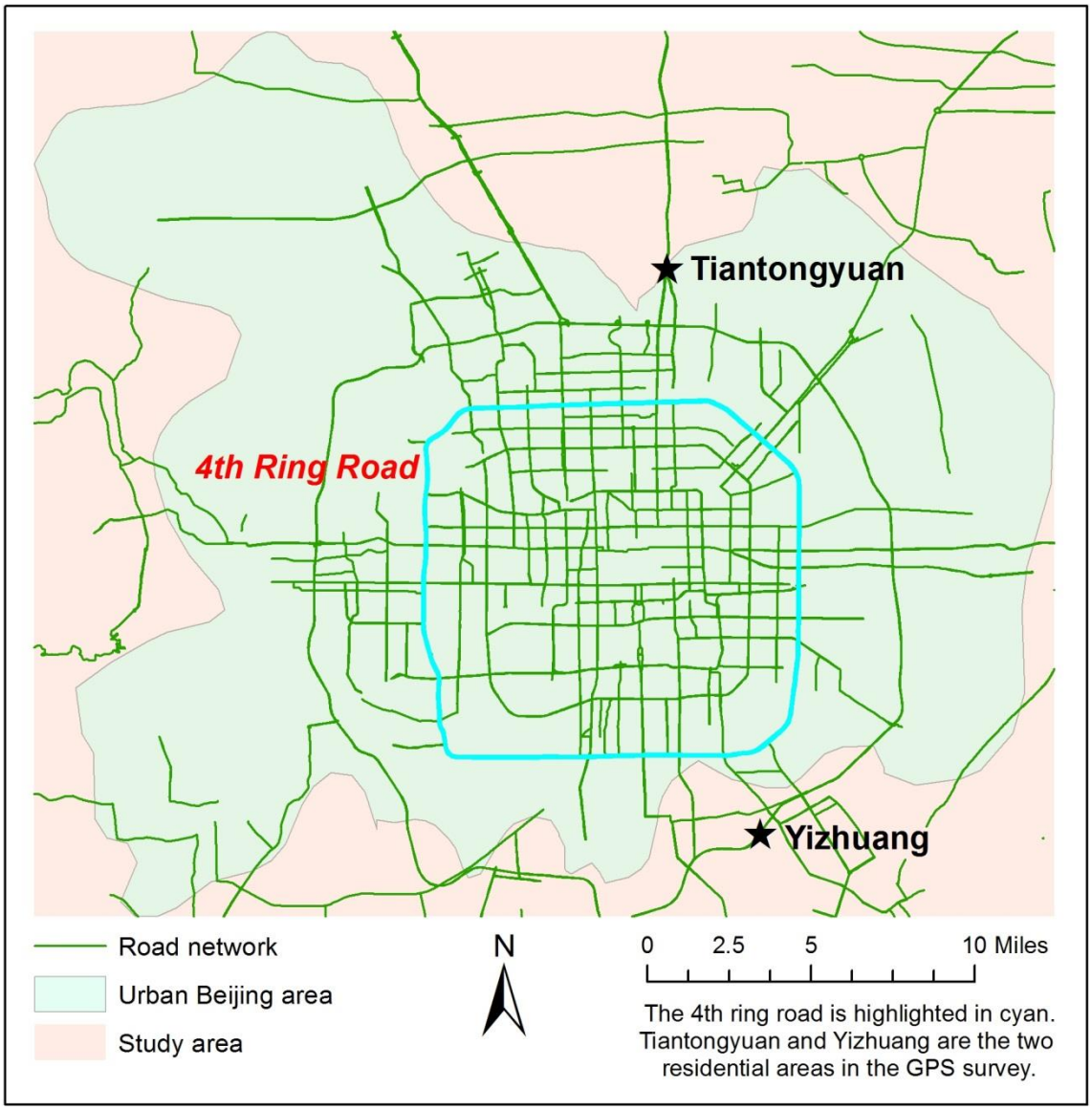

Figure 3. Road network in study area

Theoretically, the GPS positioning in this survey records the locations of the respondents every three minutes over the nine-day period. In addition to the travel diaries, each record

contains information of assigned ID, latitude and longitude, date, time. However, the data quality varies significantly across individuals not only because of the occasional GPS signal blocking and the existence of noises, but also due to the varying level of completion of participants. Paths collected from some participants were not satisfying or incomplete due to difficulties such as loss of battery power and the inconvenience of carrying a GPS device. However, since this work addresses the GPS positionings in an aggregated way, the individual level difference in data quality is not an important issue.

A sample GPS path from one participant is presented in Figure 4. The one in the left draws the path in a horizontal view in which each point is a GPS positioning record while different colors indicate different dates. The one in the right draws the trajectory using time as the vertical axis. This vertical view is probably not good at showing the exact locations but is a great tool to illustrate temporal dynamics. After aggregating trajectories from the one hundred participants and removing a few outliers, an analytical GPS dataset was produced that consists of 60,349 positionings from weekdays and 25,922 positionings from weekends. Most human activities are proved to be repeated on daily basis, but patterns in them are very different between weekdays and weekends due to regular work schedule. In this study, I excluded the GPS points from weekends, and considered only weekday patterns from ten working days. Weekend patterns will be addressed in the future work.

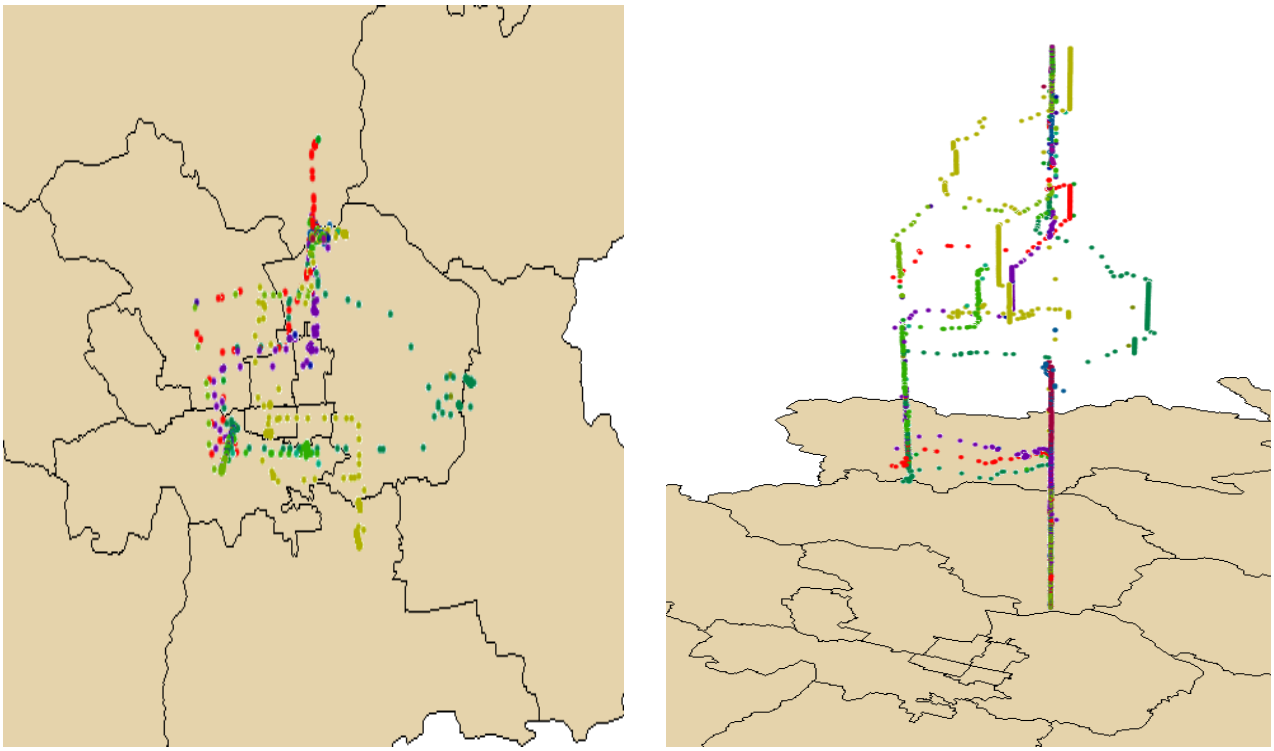

Figure 4. A sample GPS path in horizontal and vertical views

3.3 Land use classification

In dasymetric mapping studies, land use is the most commonly used ancillary dataset. It has been proved to be successful as both limiting variable and relating variable to map population and crop distribution. While the Beijing Municipal Bureau of Land and Resources has created public land use maps for a few districts, many are still unavailable. In addition, the available maps are in paper versions that are difficult to calibrate due to the lack of metadata about coordinate system and projection. More importantly, the land use classification systems are not consistent across maps from different districts. I thus decided to derive a classification of land use categories for Beijing through satellite imageries.

The satellite imagery used in this project to produce land use categories was selected from the Landsat ETM+ sensor because of its high resolution and wide availability (NASA Landsat Program, 2006, Landsat ETM+ scene 250-708, Surface Reflectance, GLCF, College Park, 09/06/2006). As reviewed in Chapter 2, the cell size of mobile phone service network usually ranges from 2000 square meters to 100 square kilometers. A 30-meter spatial resolution of the Landsat ETM+ imagery (band1-5, band 7) is thus capable to produce land use categories in a precision that has the potential to refine the mobile positionings.

The high density of tall buildings in the Beijing urban area makes the land use classification a challenging part of this work. In current remote sensing classification applications, traditional pattern recognition classifiers are still the most widely used method (Li et al. 2005). These approaches include supervised classifications (for example, K-Nearest Neighbor, Decision Tree Classifier and Maximum Likelihood) and unsupervised classifiers (for example, K-means and ISODATA). However, both types of classifiers are not robust enough because of the variation of spatial resolution across imageries and the existence of the "same spectrum with different objects" and the "same objects with different spectrum" phenomena. Other methods, such as neural network classifiers and geographic data integration, have thus been introduced for the classification of remotely sensed imageries. In land use classification for urban areas, a nonparametric method (e.g., neural networks) has been proved to be more robust in training site heterogeneity and results in a higher visual accuracy than a parametric classifier (e.g., maximum likelihood) (Paola and Schowengerdt 1995). However, a big drawback of this method is the large training times necessary for mean square error minimization.

In this work, a two-step process was employed in the land use classification for Beijing urban area. The first step used a supervised maximum likelihood classifier, in which I had three

different people working on training subsets to minimize the personal bias. Previous studies in Beijing urban area suggested a seven-category land use system which included high density built-up land, medium/low density built-up land, water body, cropland, orchards, shrub/brush, and forest land (He et al. 2001). Based on the suggested system and the particularity of my work, I first developed a four-class land use system: high density built-up land, low density built-up land, water body, and vegetation, for the supervised classification. The second step is a post-classification process. In this step, I divided the vegetation class into two subclasses: urban vegetation and suburban vegetation. This was simply based on the spatial relationship to the Fourth Ring Road (Figure 3): the vegetation inside the Fourth Ring Road was defined as urban vegetation; the outside part was defined as suburban vegetation. In addition, another class of roads was added to the existing land use categories. The road class was not set into the supervised classifier due to the difficulty of extracting linear features from remote sensing imageries. Instead, I overlaid an external data of road in Shape file format onto the existing land use classification raster and converted the pixels beneath the roads Shape file to road pixels.

The road type was included into the land use classification due to two major reasons: 1) travel activity holds a considerable proportion among all human daily activities, making road as an important land use category and thus necessary to be highlighted from other land use types; 2) road plays an essential role in the POI pattern analysis, which will be discussed more in the following section. As a result, the land use system eventually includes six categories: high density built-up land, low density built-up land, water body, urban vegetation, suburban vegetation, and road. This six-category system was applied as the basis of the ancillary data in the *Land-use method*.

3.4 Point of interests analysis

The POI dataset[2] contains 35,419 different cases from the Beijing area, including banks, hospitals, parks, shopping malls, schools, supermarkets, etc. In order to better organize and analyze the dataset, the first step of dealing with the POI dataset was to build a classification system for different types of POIs. I used a time-consuming but consistent process to classify the POIs as follow:

1. Build a preliminary system of five categories based on the function of the POIs, or the services provided by the POIs. They are schools, hospitals, shopping centers, office buildings, and residential communities;
2. Assign POIs to a category manually, one by one based on name. Unclear or ambiguous names were searched in the internet.
3. If there was no good match of existing categories to a POI, a new category was added to the system.

The final POI function system was built upon this process, and ended up with twenty-one categories (as shown in Table 1). The distribution of the POI types has the potential of acting as a good proxy for exploring the different functional districts within a city. However, it is not as valuable in the context of human mobility. Therefore, I reclassified and grouped the twenty-one functional types into different activity types based on the purposes of citizens going there. The seven activity types I finally have are: shopping activity, entertainment & recreational activity, working activity, commuting activity, house activity, eating activity and other activities. For example, it is very likely that people go to tourism sites and cinemas for entertainment or recreational activities, to Government departments and office buildings for working activities.

[2] Data from the Geosoft Lab, Peking University

However, some of the POI types were included into “other activities” because it is either difficult to tell people’s purposes of going there, or not clear of what the functions or services the POIs provide.

Table 1. Summary of POI classification

POI Activity Type		POI Function/Service Type	
1	Shopping	1	Shop
2	Entertaining & Recreational	2	Tourism site
		3	Cinema & Theater
		4	Entertainment
		5	Temple
3	Working	6	Government
		7	School
		8	Office building
		9	Research institute
4	Commuting	10	Transportation service
		11	Traffic Node
5	House activities	12	Residential
6	Eating	13	Restaurant
		14	Bakery
7	Other activities	15	Hotel
		16	Library

		17	Hospital
		18	Parking lot
		19	Bank & ATM
		20	Post office
		21	Others

Once the POI activity system was created, a connection between POIs and GPS points was expected to be built to act as the related variable in the *POI method*. The assumption here is that the presence of a citizen at a certain location and a certain time is activity-oriented, that is, every GPS point location is related to the activity types of the neighborhood POIs. An intuitive way to build this connection is to assign the geographically nearest POI to each of the GPS points. Therefore, I spatially joined all of the GPS points to their nearest POIs except for the commuting activity. As mentioned in the Section 3.3, the land use type of road plays a significant role in the POI pattern analysis. Specifically, all GPS points that fell on road pixels were tagged with commuting activity, instead of spatially joining a GPS point to a commuting POI. This was because the continuous road pixels are more likely to match the commuting activity accurately to GPS point than the discrete POIs do. As a result, each of the GPS points was tagged with an activity stamp, which could be interpreted as that "the person who traveled to this GPS point was doing the corresponding activity at that location".

The relationship between number of GPS points and associated POI activity types serves as a potential related variable for the idea of dasymetric mapping. However, it only pertains to point patterns, but doesn't provide any information about areal units that is necessary for

dasymetric mapping. Thus I modeled the 50.5km by 43.5km study area by dividing it into grids of cell size of 500 by 500 meters, and then overlaid the POIs on this graticule network (see Figure 5). The numbers of POIs with associated activity types were recorded for each cell in the grids. To characterize the cells based on their POI information, I first computed the proportion of POIs for each activity types within each cell, and used the most frequent POI activity type as the tag for this cell. For example, if the POIs in a 500x500 m^2 cell are 10% shopping, 25% eating, and 65% working activities, this cell would be characterized as a working cell because working is the dominant activity. However, a big drawback of this method is that all POIs other than the ones with the most frequent activity category were dropped within each cell. A considerable portion of useful information (dropped POIs) made no contribution to the dasymetric mapping process.

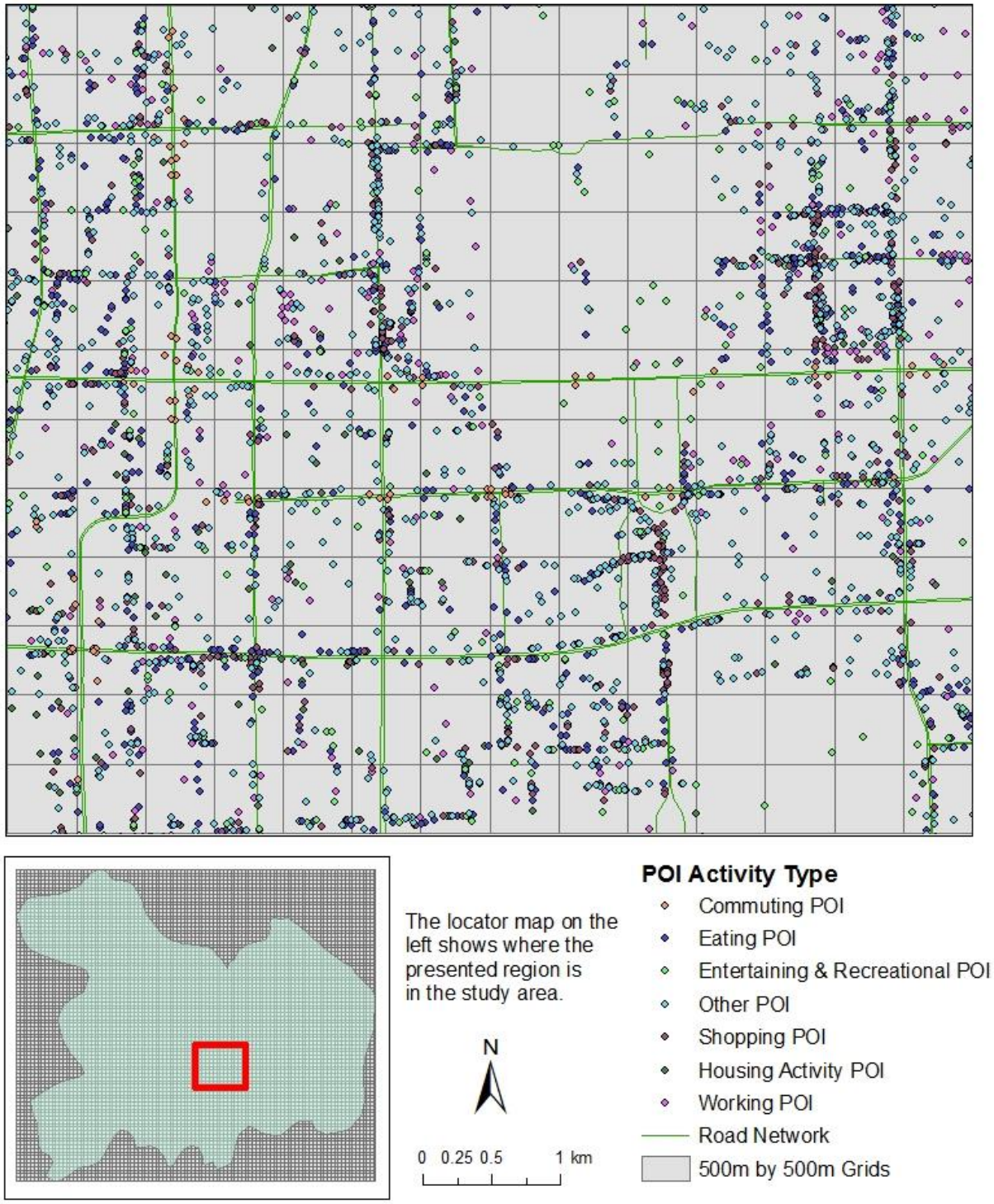

Figure 5. Different POI activity types over 500 by 500 meter grids

To avoid of dropping useful information, I used a second method that applied the ideas from cluster analysis to take all of the POIs into consideration. Again, the numbers of different POI activity types were recorded for each geographical cell, but rather than making use of only the most frequent POI, I employed a K-means algorithm to assign the closest cluster center to each cell. For example, the cell with 10% shopping, 25% eating, and 65% working activities might be placed into the cluster type of 10% shopping, 30% eating, 50% working and 10% other activities. The key question of this method is that how many clusters should be kept. In most cases, more clusters provide more accurate classification, but also result in a higher complexity of the system. A typical way to help determine the number of clusters is to seek for the smallest within-cluster variance, and the largest between-group variance. Thus I plotted the within-group sum of squares against the number of clusters, and looked for the "elbow" on the plot to choose the number of clusters.

To evaluate the fitness of clustering, I applied the Silhouette technique to test how well each cell lies within its cluster. Silhouette value varies between -1 and 1, and close to 1 means the element is well matched to its own cluster but badly matched to its neighbor cluster. Thus the greater Silhouette value the better the clustering is. I calculated an average Silhouette width for each clustering to have a better interpretation of the number of clusters picked from the within-group sum of squares method. I also examined the number of negative Silhouette values for each clustering, the ones with less negative Silhouette values are more robust because they have less chance to place a cell into an inappropriate cluster.

Given a combination of POI activity types, each cell in space was grouped in to a cluster. A map of the study area modeled by POI based clusters could thus be produced. This map is similar to the activity-aware map developed by Phithakkitnukoon et al (2010) that helps identify

the probable human activities associated with a specific region. It presents the probable services and activities that citizens could go for in each 500x500 m^2 cell, providing clues to the distribution of the functional districts in the study area. The map results will be introduced in greater details in Chapter 4.

3.5 Daily activity dynamics

From the previous processes, I created two maps that could be used in the mapping. One is the land use classification map, the other is the activity-aware map. Instead of performing a spatial joint to connect with the GPS points, I overlaid all of the GPS points on the two maps, in an attempt to build the relationship between human mobility and the ancillary information (land use categories, or POI cluster types). Since the number of GPS points from one user at a certain location doesn't necessarily illustrate the time he/she stayed there, I further explored how much time a person spent at each position of his/her GPS trajectory. The relationship between the accumulative staying time instead of the number of GPS points, and the associated land use category or POI activity cluster type was used to map the call records.

The movements derived from one user's GPS trajectory is a temporal sequence of locations. Although the time of the sequence always increased at a monotone rate, the way to get accumulative staying time at each location would not be simply calculating the time difference between adjacent positions in a trajectory, due to the existence of signal noises and floating points in GPS data. Let P_{GPS} denote one user's GPS-based path such that

$$P_{GPS} = \{< x_{t_i}, y_{t_i} > | \, i = 0,1,...,n\}$$

where $< x_{t_i}, y_{t_i} >$ is a pair of coordinates for this user at time t_i . $< x_{t_0}, y_{t_0} >$ and

$< x_{t_n}, y_{t_n} >$ are the starting point and ending point of this path, respectively. The algorithm I used to derive staying time for each location involved two spatial thresholds s_1, s_2 and a temporal threshold t. s_1 is a threshold used to get rid of signal noises, and s_2 is a threshold to remove the impacts of floating points where they such that $s_1 < s_2$. The algorithm dealt with three different scenarios:

Scenario1: *If the displacement between* $< x_{t_i}, y_{t_i} >$ *and* $< x_{t_{i+1}}, y_{t_{i+1}} >$ *is greater than the noise threshold* s_1*, but less than the floating threshold* s_2*, they are in a movement. Cumulate the time difference* $\Delta t = t_{i+1} - t_i$ *to the earlier position* $< x_{t_i}, y_{t_i} >$*, and move to the later position* $< x_{t_{i+1}}, y_{t_{i+1}} >$*;*

Scenario2: *If the displacement between* $< x_{t_i}, y_{t_i} >$ *and* $< x_{t_{i+1}}, y_{t_{i+1}} >$ *is smaller than the noise threshold* s_1*,* $< x_{t_{i+1}}, y_{t_{i+1}} >$ *is regarded as a signal noise. Cumulate the time difference* $\Delta t = t_{i+1} - t_i$ *to the earlier position* $< x_{t_i}, y_{t_i} >$*, and stay on it;*

Scenario3: *If the displacement between* $< x_{t_i}, y_{t_i} >$ *and* $< x_{t_{i+1}}, y_{t_{i+1}} >$ *is greater than the floating threshold* s_2*, look at the time difference* $\Delta t = t_{i+1} - t_i$*:*

a. *If* Δt *is greater than the temporal threshold* t*, there are some data missing during this* Δt*. Don't accumulate staying time, and directly move to the later position* $< x_{t_{i+1}}, y_{t_{i+1}} >$*;*

b. *If* Δt *is less than the temporal threshold* t*,* $< x_{t_{i+1}}, y_{t_{i+1}} >$ *is a floating point. Don't accumulate staying time, ignore* $< x_{t_{i+1}}, y_{t_{i+1}} >$ *and move the next*

$$position < x_{t_{i+2}}, y_{t_{i+2}} >.$$

The high temporal resolution of GPS data is an important reason that GPS devices are widely used in time geography research. Since locations of individual can be tracked clearly over time as a somehow continuous trajectory, time difference is able to be considered in analyzing human daily movement patterns. In existing literatures, GPS positions had been divided into twenty four 1-hour bins/eight 3-hour bins throughout a day, or been regrouped based on human social activities into, for example, a before-working-period, a during-working-period, and an after-working-period.

In this study, I built the time differential relationship between land use category and the associated staying time, as well as that between POI activity cluster type and the associated staying time. The temporal sequential relationships were built upon eight 3-hour bins as shown in Figure 6, providing the potential to better estimate the actual locations of mobile phone positionings within the service cells. That is, rather than one fixed related variable, eight time-specific relationships could be applied in the mapping of mobile positionings according to when the mobile telecommunication was established.

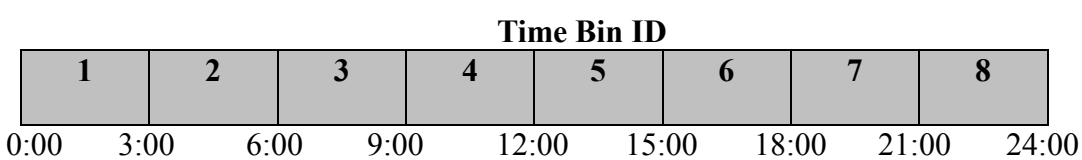

Figure 6. Alignment of the 3-hour time bins

Once the staying time at every GPS location was achieved, the human activity intensity λ_{ij} in this work was defined for both *Land-use method* and *POI method* as:

$$\lambda_{ij} = T_{ij} / A_i$$

Where i is the land use category ID in the *Land-use method*, or the POI cluster type ID in the *POI method*. j is the time bin ID from Figure 6. A_i is total area of land use i or POI cluster i, T_{ij} is the accumulative GPS staying time during time bin j on all areas characterized as land use i or POI cluster i. The activity intensity λ_{ij} is varying on both space (i) and time (j), and the unit of it is minutes per square kilometers, for example, 100 min/km^2 means that people stay for 100 minutes in every one square kilometers space at this location.

3.6 Cellular service network simulation

A known mobile phone service network is helpful to validate how much improvement on spatial accuracy of mobile positionings this method can gain. However, the spatial distribution of the mobile phone antennas (base towers) is not available from the study area. An alternative, though not the best, way is to simulate the distribution of antennas upon other related information or indicators.

A number of existing literatures has suggested that the mobile phone antennas are set up according to the local population, the distribution of antennas is very much correlated to the underlying population distribution (Kang et al. 2012). Because of the limited capacity of a single antenna, it is reasonable that population-dense areas require more antennas to serve a larger amount of telecommunication than sparse areas do. I thus decided to use population as an indicator (LandScan 2008 population dataset) to generate the coordinates of the mobile phone base towers in the study area. When creating the service network from the base towers, the principles of how the antennas work were applied. The Cell-ID based network normally delivers

the data from a telecommunication activity to and records at its nearest antenna, and then transmits out to its destination. This working principle perfectly matches the method of creating thiesson polygons (Voronoi diagram) from their centroids. The simulation procedures I performed were:

1. Prepare a population distribution in raster format for the study area. The data comes from LandScan 2008 with a 30X30 second resolution ;
2. Estimate the total number of antennas in the study area. The study area is 50.5km by 43.5km, and the average cell size of mobile phone network is approximately 2km (Ahas et al. 2007a). I thus estimated a number of 550 antennas using these two parameters;
3. Generate the coordinates of the 550 antennas across the study area based on the underlying population distribution raster;
4. Create the cellular service network. I used the 550 antennas as centroids to create a theissen polygon network clipped by the study area (Figure 7). Each theissen polygon is a service cell.

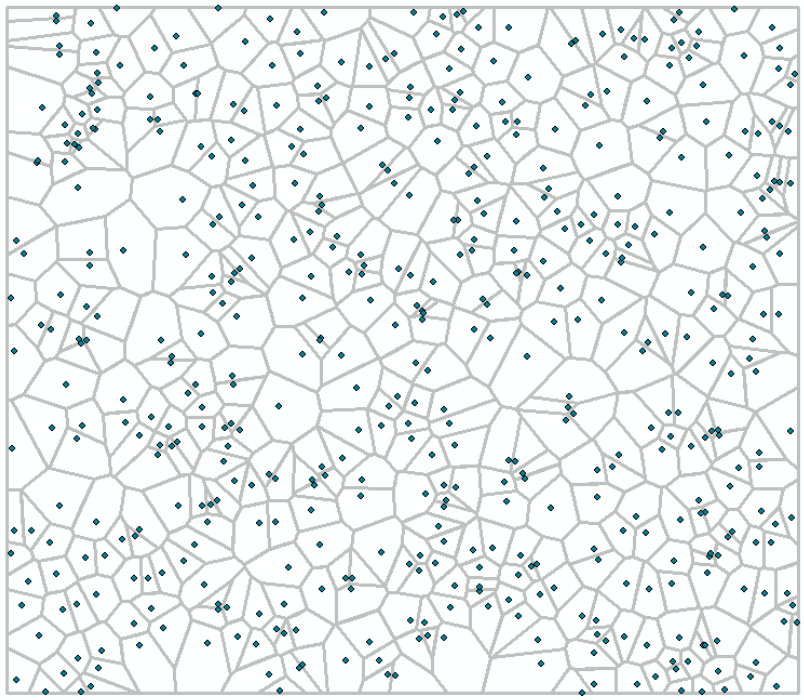

Figure 7. Cellular service network and antennas in the study area

When the mobile phone service network was established, the next step is to simulate the CDRs from GPS-based trajectories. I extracted the time stamp (when) and the antenna (where) if the individual established a telecommunication with another person. Often in studies of human mobility the distribution of the time intervals between consecutive activities, such as e-mail communications, phone calls and web browsing, in general follows the so-called power-law (Barabasi 2005). In other words, individual's call activity has internal patterns, which can be modeled and predicted. Thus, I iteratively predict the time of individual's next call activity based on this law, and generate his/her complete CDRs. Different from the GPS-based path P_{GPS}, the individual's CDR-based path P_{CDR} can be defined as,

$$P_{CDR} = \{< x'_{t'_j}, y'_{t'_j} > | \; j = 0,1,...,m\}$$

In temporal dimension, the one user's path of CDRs is derived from his/her GPS-based path, thus has a lower resolution

$$\{t'_j \mid j = 0,1,...,m\} \subset \{t_i \mid i = 0,1,...,n\}$$

where m, n ($m \leq n$) are the number of points in the GPS-based path and the CDR-based path respectively, and t'_j is the time when the antenna routing the subscriber's telecommunication. In spatial dimension, coordinates of the antenna are taken as subscriber's position, thus we denote subscriber's location at time t'_j as

$$x'_{t'_j} = x_{t'_j} + \Delta x$$

$$y'_{t'_j} = y_{t'_j} + \Delta y$$

where Δx and Δy are errors determined by size of the service area of each antenna.

3.7 Refinement in call records distribution

In point pattern analysis studies, the realization of a Poisson random variable is widely used to model the possibility of a point falling into a sub-region B in the study area. Since a Poisson distribution is controlled by a single parameter, intensity λ, the expected possibility P of a point falling into a sub-region B is,

$$P = \lambda \times area(B)$$

In a Cell-ID based mobile phone service network, the location recorded for each call activity is referred to the coordinates of the nearest antenna. It is not possible to know where exactly the call activity was made. That is, intensity λ does not vary spatially within a service cell. This type of point pattern is called a uniform Poisson Point Process or CSR (Complete

Spatial Randomness).

The methodology of my work is to break the rules of CSR, and to make λ vary on the different land use categories or the POI activity cluster types, not only spatially, but also temporally. In this study, the activity intensity λ_{ij} is actually the previously built relationships between accumulative staying time and land use categories/POI activity cluster types. The spatial variation of λ_{ij} is associated with the partition of the land use classification map and the POI activity clustering map. The temporal variation comes from the eight 3-hour bins.

In the *Land-use method*, activity intensity λ_{ij} is decided by land use category and time. If the area of a sub-region is known, the expected probability (activity magnitude) of a point falling into the sub-region could be achieved through multiplying the intensity and area. In this work I am estimating the location of a call event within a mobile phone service cell, the probability needs thus to be normalized by each service cell. Since every sub-region (pixel) has the same size in the land use raster, the normalized probability surface in the *Land-use method* could be produced through each pixel:

$$P_k = \lambda_{ijk} / \sum_i \lambda_{ij} \times N_i$$

where P_k is the normalized probability of a call event falling into pixel k. λ_{ij} is the activity intensity where i is the land use type pixel k belongs to, and j is the time bin ID when the call event established. N_i is the number of pixels in land use type i within the mobile phone service cell.

Similarly in the *POI method*, activity intensity λ_{ij} is decided by POI cluster type and time. However, the sub-regions are no longer the raster pixels, but the parcels produced by

intersecting the 500 by 500 meters grids and the cellular service network. Since the sizes of sub-regions (parcels) are not unchanging like the raster pixel size, the normalized probability surface in the *POI method* is defined as:

$$P_k = \lambda_{ijk} \times A_k \,/\, \sum_l \lambda_{ijl} \times A_l$$

where P_k is the normalized probability of a call event falling into parcel k, λ_{ijk} and A_k are the activity intensity and area of parcel k, respectively. λ_{ijl} and A_l are the activity intensity and area of the l th parcel in the service cell which parcel k belongs to.

Chapter 4 Results and Analysis

4.1 Land-use based mapping

As discussed in Chapter 3, a supervised classification was used to get different land use categories in the study area. After combining the training subsets from three people's work, a preliminary classification map was produced through the Maximum Likelihood algorithm in ENVI. Figure 8 shows the classification results in the central part of the study area. Four land use categories were achieved in this output: high density built-up land, low density built-up land, vegetation, and water bodies. In this transitional map result, high density built-up land occupies the majority of the study area. This might impact the effectiveness of distinguishing the space, especially in central urban areas where the human activity intensity is high. In the post-classification process, the vegetation was divided into urban vegetation and suburban vegetation based upon its spatial relationship to the Fourth Ring Road. A special land use category, road, was also included into the existing classification due to its importance in the activity pattern analysis. Figure 9 presents the distribution of six categories after the post-classification at the same extent of Figure 8.

With the inclusion of two more categories after the post-classification, the map results have more land use diversity than the initial supervised classification results do. Although high density urban land is still the largest proportion of the study area (39.03%, Table 2), it has a lot more heterogeneity especially in areas where road network is dense. The second largest land cover is suburban vegetation that consists of some forests on northwest mountains and lots of croplands on plains in the suburbs of Beijing. However, urban vegetation is very small (1.04%), about the same size of water bodies (1.05%), suggesting a low greenery coverage in urban

Beijing. While using the staying time calculated from the GPS dataset as an indicator for human activity intensity, a summary of the relationship between human activity and the six land use categories was generated in Table 2. It is not surprising that road land use has the highest activity intensity of 501.90 min/km^2 because it has a relatively small area (10.36%) but a great amount of commuting activities. Although high density built-up land has more than half of the total GPS staying time (56.03%, which is more than twice of road), its activity intensity (324.63 min/km2) is lower than road due to its large area.

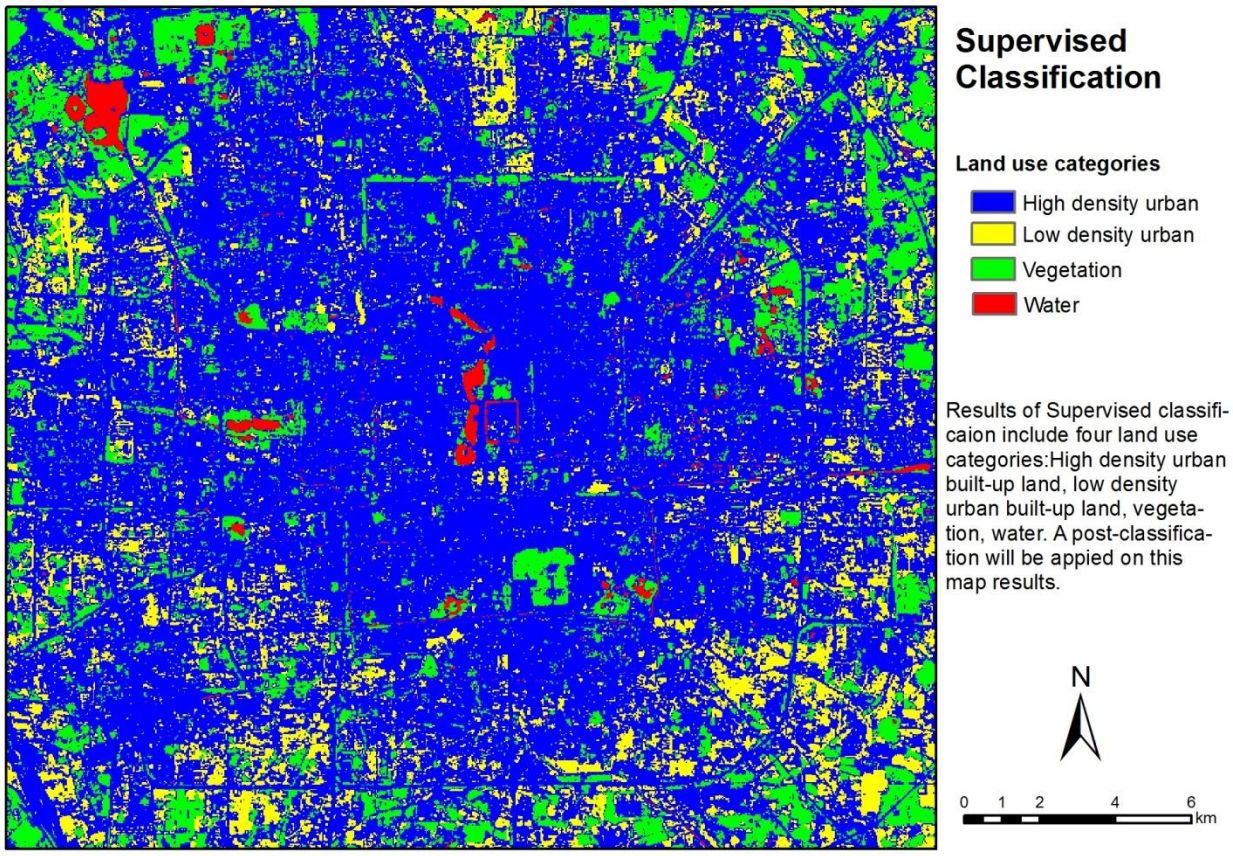

Figure 8. Supervised classification results in central study area

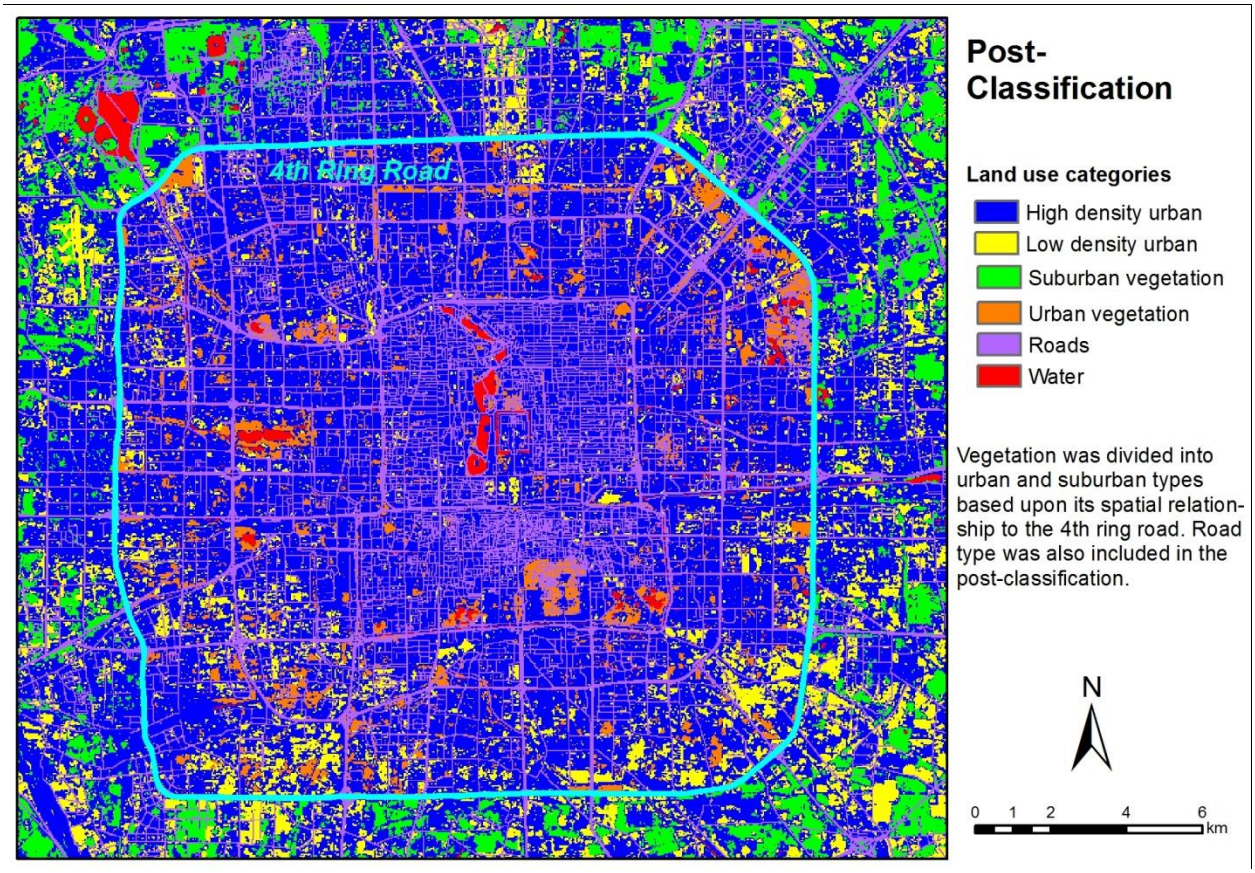

Figure 9. Post-classification results in central study area

Table 2. Land use classification results (one hundred individuals over five weekdays)

Land Use Category	Number of Pixels	Area (km^2)	Percent Area	GPS Staying Time (min)	Percent GPS Staying Time	Activity Intensity (min/km^2)
High density urban	946,092	851.48	39.03%	276,419	56.03%	324.63
Low density urban	321,390	289.25	13.26%	80,590	16.34%	278.62
Suburban vegetation	854,986	769.49	35.27%	16,554	3.36%	21.51
Urban vegetation	25,159	22.64	1.04%	4,578	0.93%	202.18
Road	251,108	226.00	10.36%	113,427	22.99%	501.90
Water body	25,442	22.90	1.05%	1,768	0.36%	77.21

The two panels in Figure 10 plot the human daily activity dynamics against time bin ID by land use category. Time bins from 1 to 8 goes from 0:00-3:00 to 21:00-24:00, which refers to the alignment in Figure 6. Each color in the plots indicates one land use category. The bottom panel shows the proportion of accumulative staying time (activity magnitude) for each land use category among all categories. An overall pattern here is that the difference in activity magnitude across land use categories is smaller in the daytime (curves tend to converge), but becomes larger at night (curves tend to diverge). Most activities appeared on high density built-up land, road, and low density built-up land while the other three land use categories only have a small proportion of the total staying time. High density built-up land has the largest activity magnitude over all time bins, and reaches its peak at 3am to 6am (Time bin 2) which is the sleeping time in regular schedules. This suggests that most people have their homes at high density built-up areas. On the other hand, low density built-up land has relatively high activity magnitude during working hours from 9am-3pm (Time bin 4 and 5), indicating a lot of people working at this land use type. As expected, the activity magnitude peak of road is at 6am to 9am (Time bin 3) because a considerable amount of commuting activities appears during this time period. The activity intensities in the upper panel were achieved through normalizing the GPS staying time by the total area of the corresponding land use. Almost all land use categories have a 'inverse U' shaped curve, suggesting a high activity intensity in the daytime but drops to a low level at night. Road has the highest activity intensity over almost all times because of its small area as mentioned.

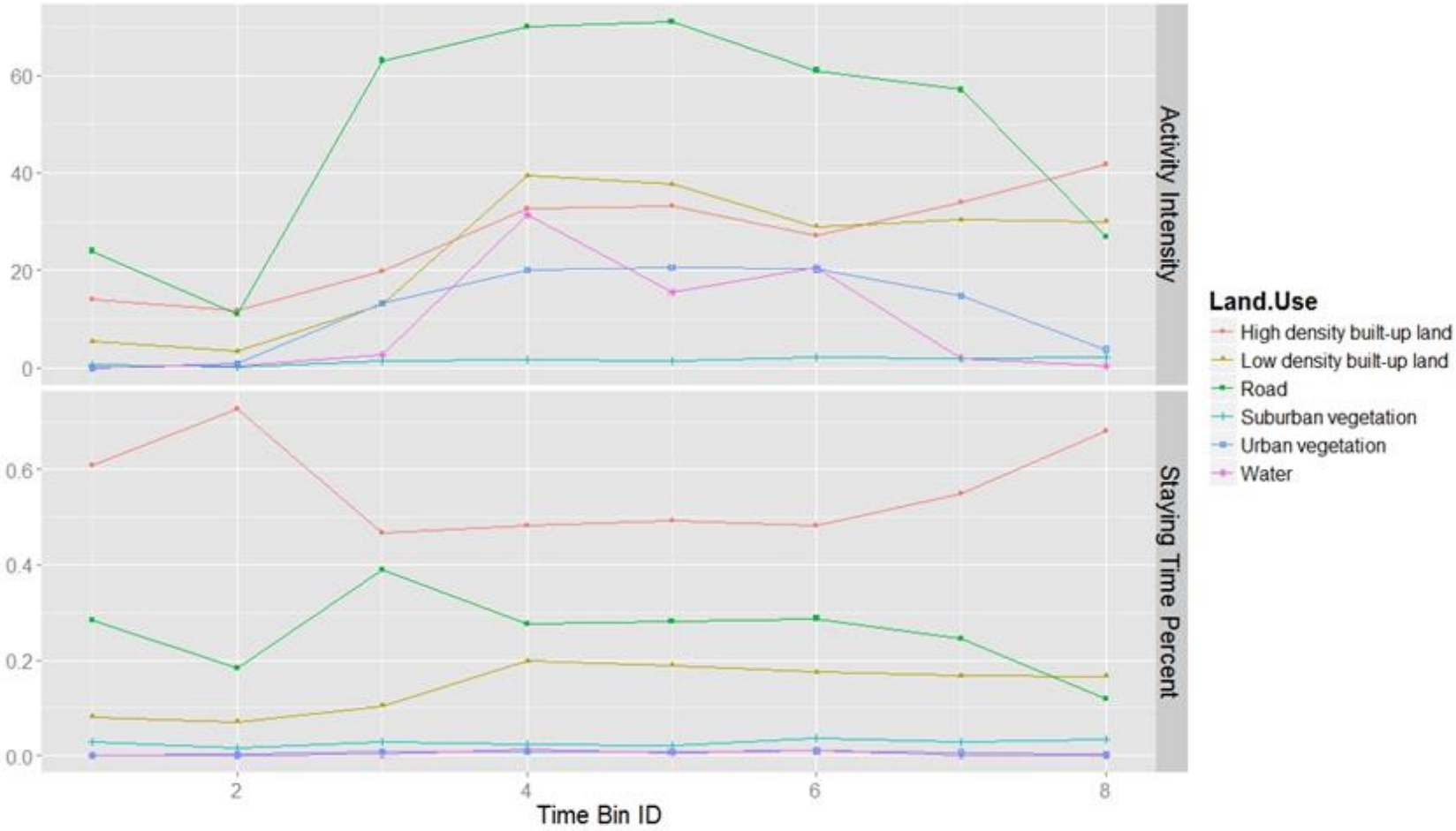

Figure 10. Daily activity dynamics by land use category

Figure 11 through Figure 14 present the results of *Land-use method* at a sample extent. The presented area is selected from central north Beijing because it has enough land use diversity and relatively high activity intensity. Figure 11 maps the land use classification to provide background information for this area. Figure 12 is the density/intensity of human activity at Time bin 3 (6am to 9am). Road pixels pop out with the highest intensity of 63.00 minutes per square kilometers, in accordance with the results from Figure 10. The cool colors with low activity intensity include water, and suburban vegetation.

While overlaid activity intensity raster on the cellular service network, I summed the activity magnitude to each mobile phone service cell. Figure 13 shows the sum activity magnitude by mobile phone service cells at 6am to 9am. Larger cells tend to have higher sum activity magnitude because they have more pixels than small cells do. Figure 14 is the resulting heat map of *Land-use method* that shows normalized activity magnitude by mobile phone service cells for each pixel in the study area. This map is also regarded as the normalized probability surface of call events in which pixel values sum up to one for every cell. The more details in land use a cell has, the more variation in probability it will have. More importantly, a pixel with high activity intensity doesn't necessarily have a high probability of a call event falling into it. For example, the pixels in the big mobile phone service cell at the right bottom of Figure 13 with the highest sum activity magnitude have almost the lowest normalized probabilities in Figure 14. In traditional mobile studies, CDRs can only be located to the centroid of its service cell, or be randomly assigned to a location within the cell. With the dasymetric mapping results from this work, locations of CDRs could be estimated based upon the normalized probability surface within each cell, providing a much higher spatial accuracy.

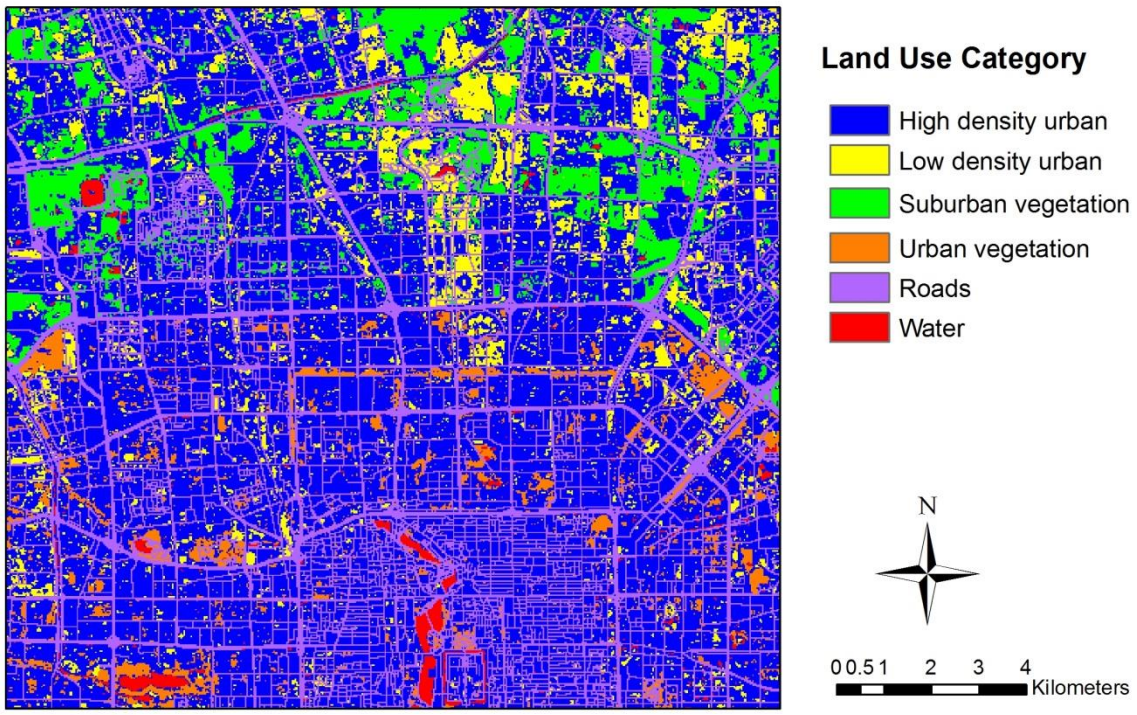

Figure 11. Distribution map of land use category

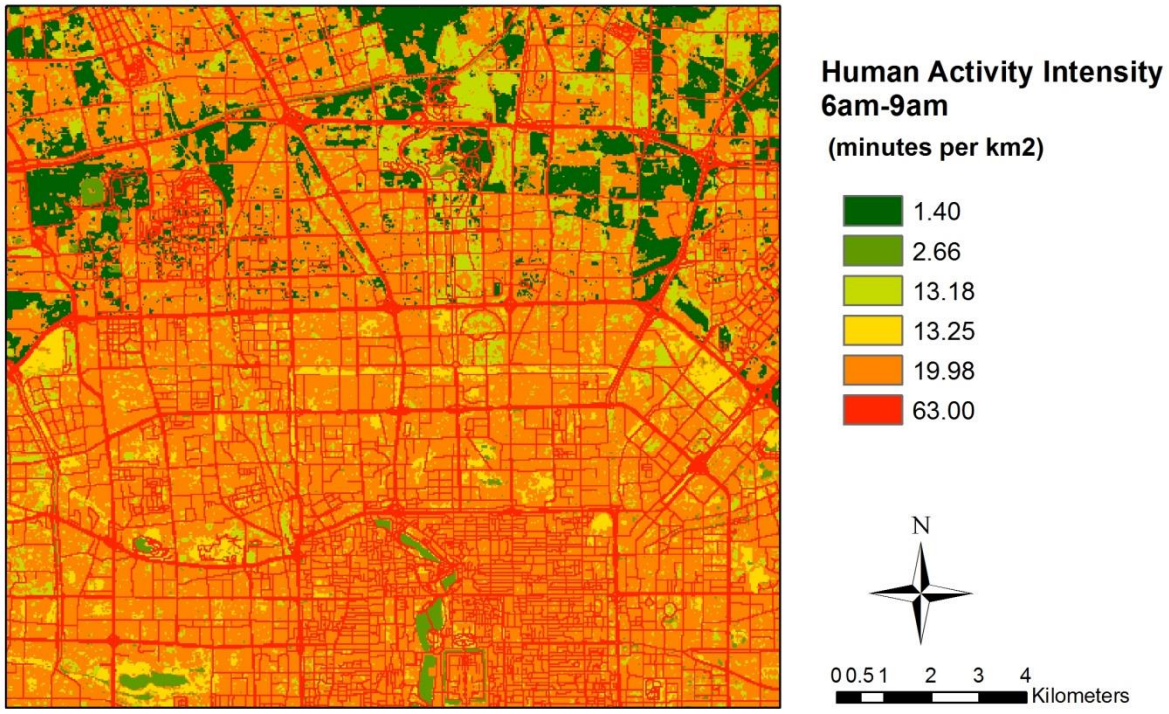

Figure 12. Intensity map of human activity by land use (one hundred individuals over five weekdays)

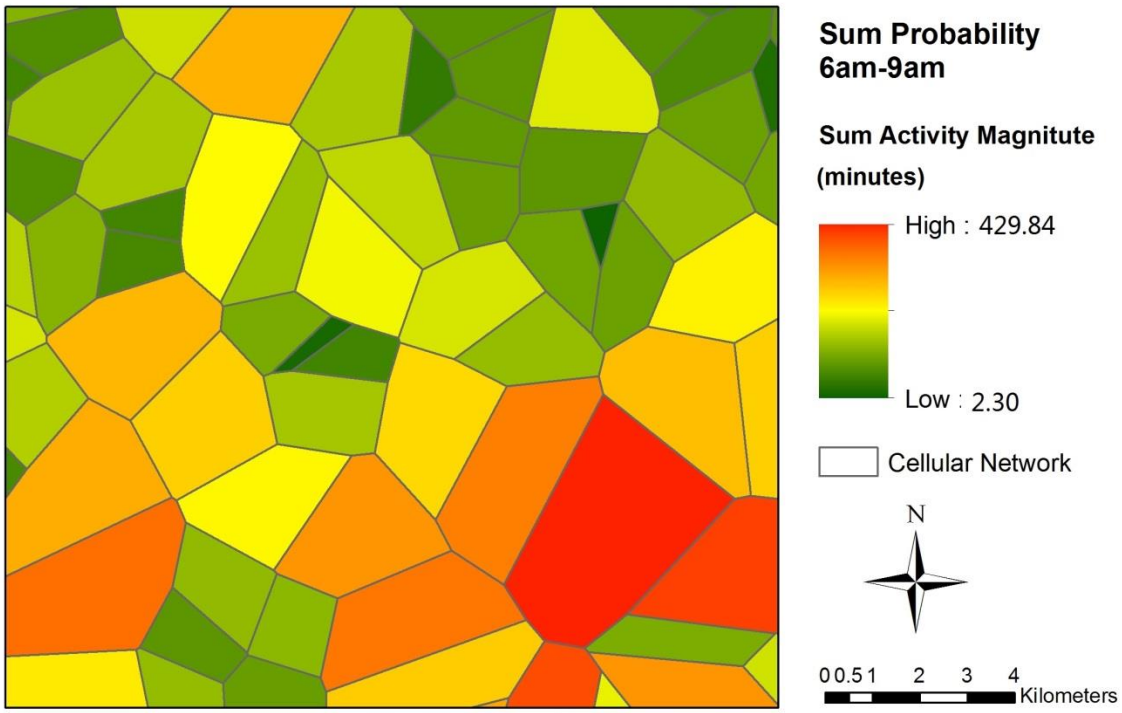

Figure 13. Sum activity magnitude by cellular service network

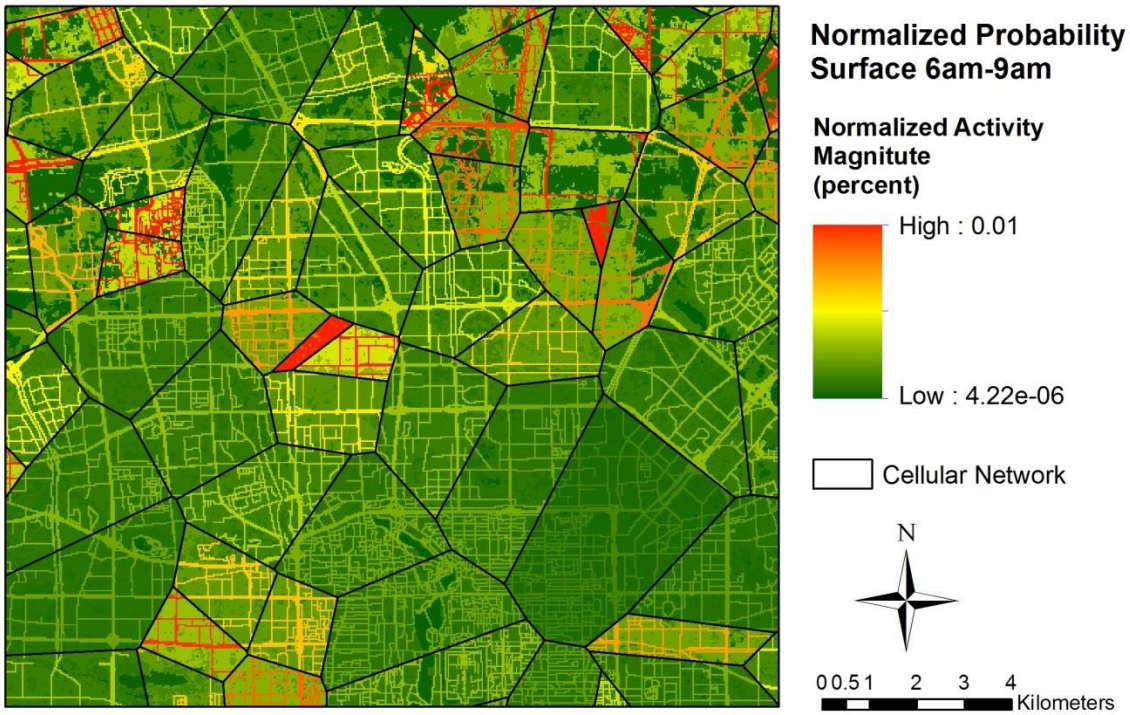

Figure 14. Normalized probability surface by land use

4.2 POI based mapping

Figure 15 plots the number of GPS points against the associated POI activity types. When compared to the number of POIs in each activity type, the greatest difference occurred in residential activities. If human movements are random walks, the two distributions should be very similar to each other. However, the proportion of GPS points associated with residential activity is much greater than the proportion of residential activity among all POIs. This suggests that people tend to stay at residential areas for a substantial time, but keep moving around across other activity areas. For example, people spend a long time at residential areas involving activities such as sleeping.

Figure 15. summary of POI activity categories

Given a combination of POI activity types for each 500 by 500 meters grid cell, I apply k-means clustering to group them in space. For better interpretation, the clustering is based upon the proportion of each activity type, rather than the occurrence. Since a considerable variation in clustering results was observed due to the random starting points, 50 clustering processes were performed for each k to avoid biased results while choosing the best number of clusters.

In Figure 16 each boxplot illustrates the within-group sum of squares of 50 clustering processes for each number of cluster centers used from K=2 to K=8. Although the within-group sum of squares goes down relatively smoothly as the number of clusters increases, we can still see a few jumps such as from K=5 to K=6, and from K=7 to K=8. Among these two jumps, the eight-center clustering seems to be more robust because it has less variation than the six-center clustering does, thus a lower chance to get an unusual clustering and outliers.

Boxplots in Figure 17 are similar to those in Figure 16, but plot the average silhouette width. Rather than the smooth decrease in within-group sum of squares, we can see an apparent ‘elbow’ at K=8 in this figure, that is, the average silhouette width goes up quickly when K is small but begins to slow down after K=8. Moreover, the Eight-center clustering has the smallest number of negative silhouette values (127 observations) among all clustering solutions, which indicates that Eight-center clustering places the smallest number of cells into inappropriate groups. All of these results suggest K=8 to be used to group the gird cells and to characterize space in the study area.

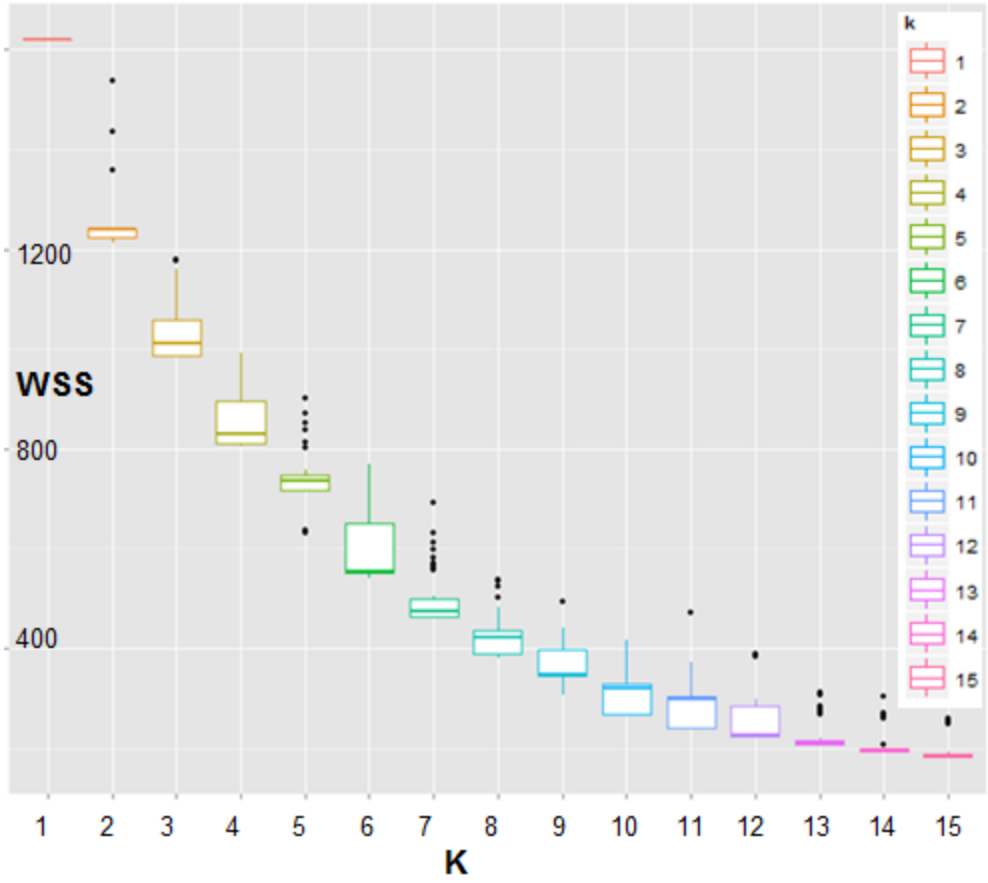

Figure 16. Boxplots of within-group sum of squares

(WSS= within-group sum squares, K= number of clusters)

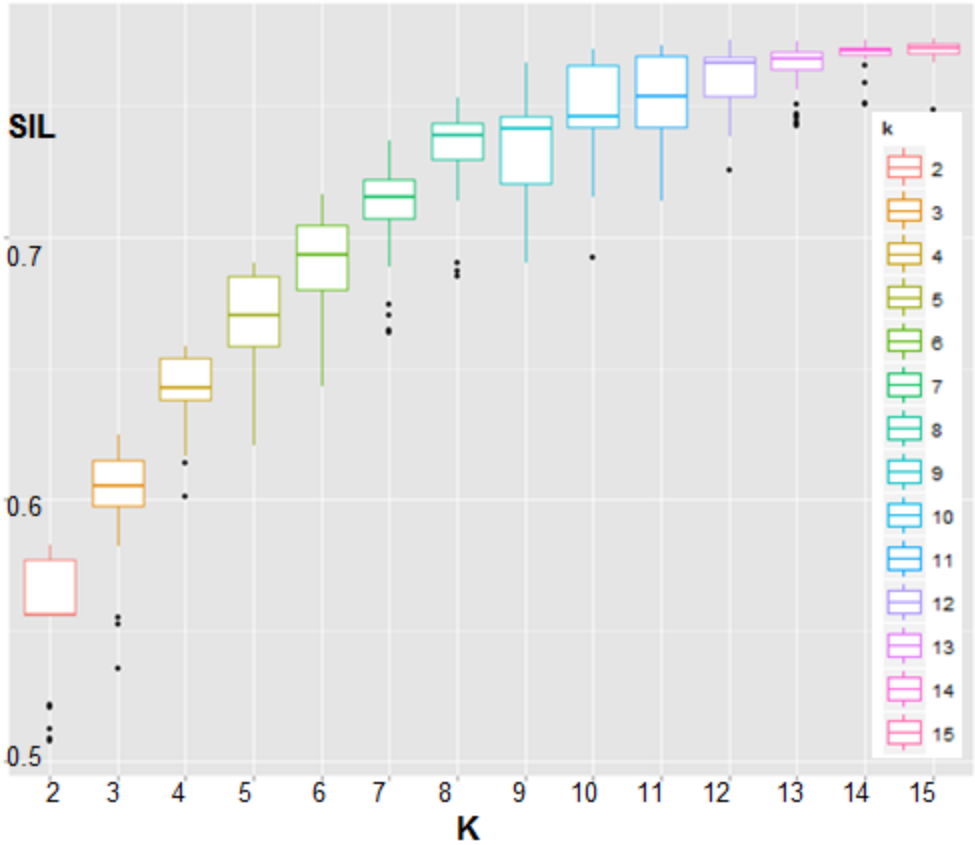

Figure 17. Boxplots of average silhouette width

(SIL= average silhouette width, K= number of clusters)

The resulting map of the clustering in the study area is presented in Figure 18 where each cell is classified into one of the eight clusters. A summary of the clusters is depicted in Table 3. To better interpret the results, I assigned an activity type to each cluster based on its center. The cluster center is a combination of the seven POI activity proportions in the following order: Shopping, Entertaining & Recreational, Working, Commuting, House activity, Eating, Other activities. For example, Cluster 5 is tagged with shopping-oriented because 93% of the POIs on it are shopping POIs while none of other activity POIs holds a proportion greater than 2%.

While combining the activity-aware map with land use and google base maps, some

interesting findings could be revealed. A large proportion of cells have no POI presence because they are rural areas such as cropland and mountains (Cluster 6). Most of the cells within the Fourth Ring Road are belong to Cluster 7 which is a mix of working and eating POIs, while most Cluster 3 (House activity-oriented) appear at suburban areas. This can help characterize human daily movements from suburban homes to working-oriented central areas in the morning, and the opposite direction during off-working hours. Another mixed activity type consists of shopping and eating POIs (Cluster 1). Most of them show up outside of the Fourth Ring Road, and usually comes with a number of entertainment POIs. These could be the places citizens likely to go after work or at weekends. It is also interesting that the shopping-oriented cells (Cluster 5) never appear in central Beijing areas, unless they are mixed with other POIs. This suggests shops are rarely located far from restaurants or working places in central Beijing.

Figure 18. Activity-aware map of study area

Table 3. Summary of POI activity based clusters

(Cluster center is a sequence of proportions in the following order: Shopping, Entertaining & Recreational, Working, Commuting, House activity, Eating, Other activities)

Cluster ID	Cluster Center	Activity Type	Number of Cells	Percent of Total Cells
1	38%-5%-4%-2%-8%-26%-16%	Mix of Shopping and Eating	689	7.8%
2	7%-2%-3%-76%-3%-5%-4%	Commuting-oriented	166	1.9%
3	3%-2%-2%-1%-85%-5%-3%	House activity-oriented	265	3.0%
4	5%-3%-2%-0%-3%-84%-4%	Eating-oriented	343	3.9%
5	93%-1%-1%-1%-1%-1%-2%	Shopping-oriented	288	3.3%
6	0%-0%-0%-0%-0%-0%-0%	None	5582	63.5%
7	9%-5%-21%-3%-6%-23%-32%	Mix of Working and Eating	1083	12.3%
8	4%-3%-3%-1%-6%-4%-80%	Other activity-oriented	371	4.2%

Similar to Figure 10, Figure 19 uses two panels to plot the human daily activity dynamics against time bin ID by POI activity cluster types. Again, time bins from 1 to 8 goes from 0:00-3:00 to 21:00-24:00 in a night-day-night sequence. The panel at the bottom plots the accumulative staying time in percentage from the GPS dataset over the 500 X 500 meters grids, which could be interpreted as the human activity magnitude by each POI cluster type. Cluster 7 reaches its peak activity magnitude during working hours (time bins 4-6, 9am-6pm) when almost all other clusters have their lowest activity magnitude, making the curve of Cluster 7 pop out from other curves. This supports the fact that Cluster 7 is the only type holding a large proportion of working POIs. It is interesting that unlike the pattern found in *Land-use method*, the difference in activity magnitude is greater in the daytime, but smaller at night across POI clusters. The variation in activity intensity across POI clusters is also smaller than that among land use categories. However, we can see an apparent opposite shape between Cluster 3 (House activity-oriented) and Cluster 7 (Mix of working and eating): the curve of Cluster 3 is U-shaped; the curve of Cluster 7 is inversely U-shaped. This illustrates people's regular schedule that high activity intensity at homes appear at night while work places have more activities in the day time. It is also interesting that Cluster 2 of commuting-oriented POIs has two activity intensity peaks at time bin 3 (6am-9am) and time bin 7 (6pm-9pm), from which we can expect a large amount human movements appearing from homes to work places at 6am-9am, and in the opposite direction at off-working hours from 6pm to 9pm.

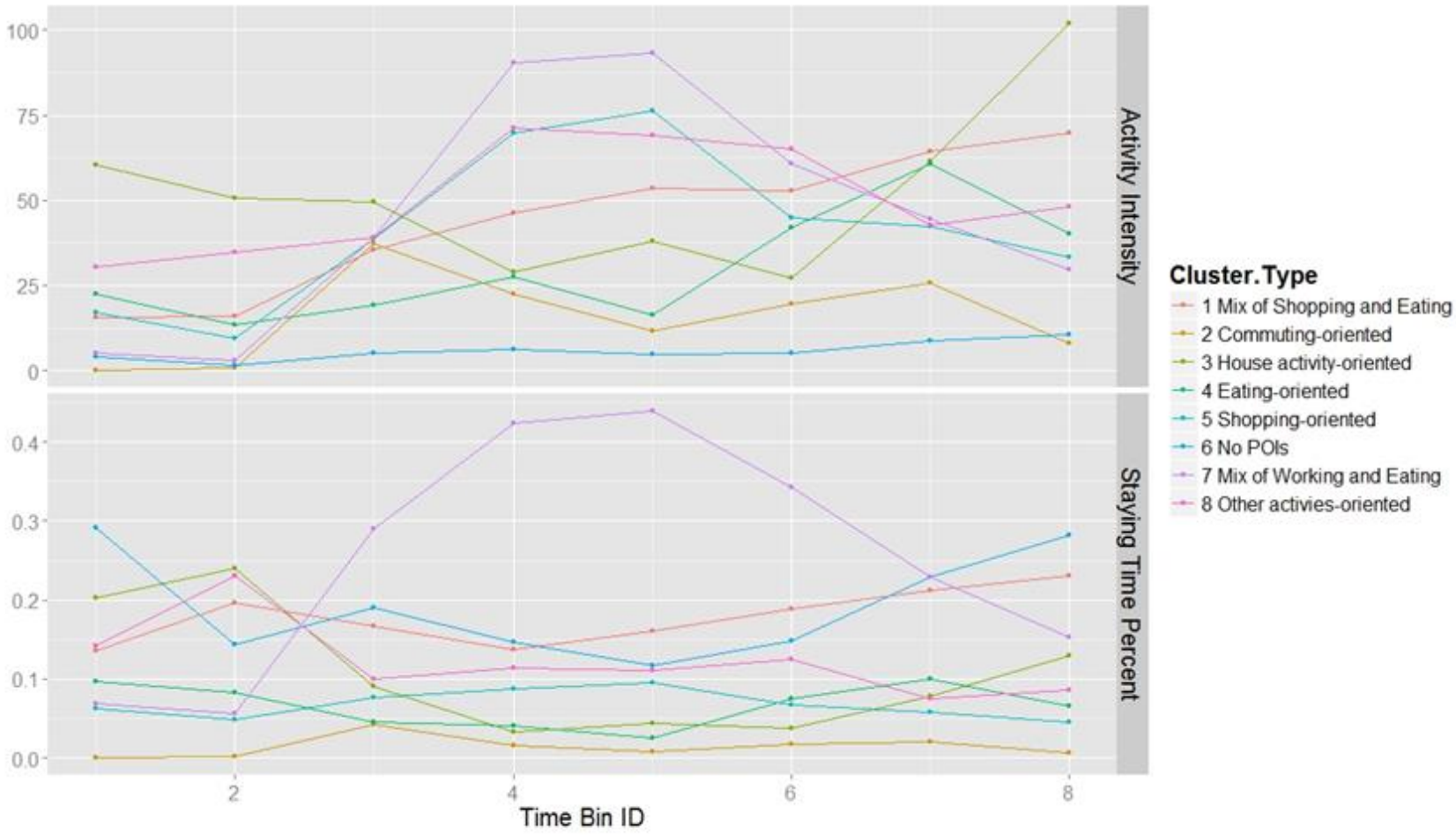

Figure 19. Daily activity dynamics by POI cluster types (one hundred individuals over five weekdays)

In the *POI method*, I was also able to create a normalized probability surface of CDRs (Figure 23), and Figure 20 through Figure 23 present the results of *POI method* processes described in Chapter 3. These map results zoomed in to the same extent of that in the *Land-use method* for comparison purposes. Figure 20 is a distribution map of the POI activity cluster types. There are a great number of cells with mix of working and eating POIs in the south area, and some cells with no POIs in the north, all other cluster types are relatively evenly distributed across the mapped area. The activity intensity heat map in Figure 21 is also from Time bin 3 (6am to 9am). House activity-oriented cells (Cluster 3) have the highest intensity of 49.66 minutes per square kilometers, and not surprisingly, the lowest intensity appears on places with no POIs (Cluster 6).

A lot of parcels were produced while clipping the grids by the cellular service network. The activity magnitudes of the parcels were calculated based on the area and activity intensity of them (Figure 22). Since this work seeks to refine the distribution of call events within each mobile phone service cell, the activity magnitude needs to be normalized. Figure 23 shows the normalized probability surface using POI based dasymetric method. Values in this map sum up to one within each mobile phone service cell, and each of values is the probability of a call event falling into the associated parcel.

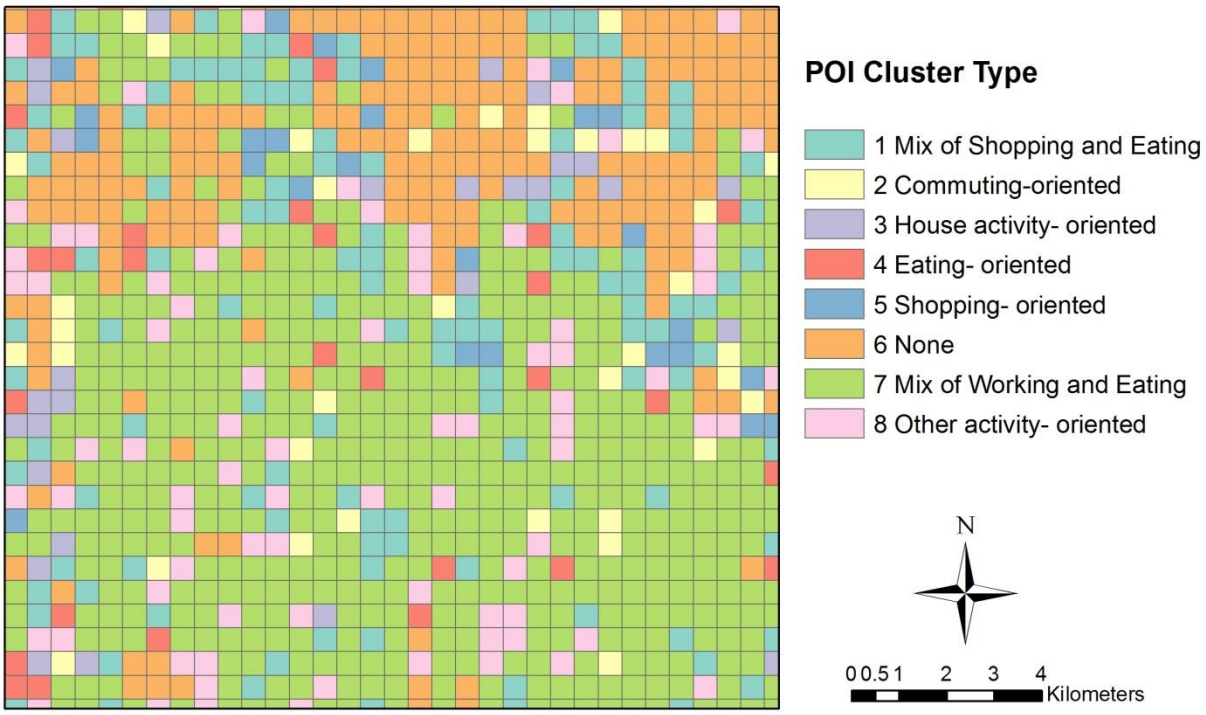

Figure 20. Distribution map of POI cluster types

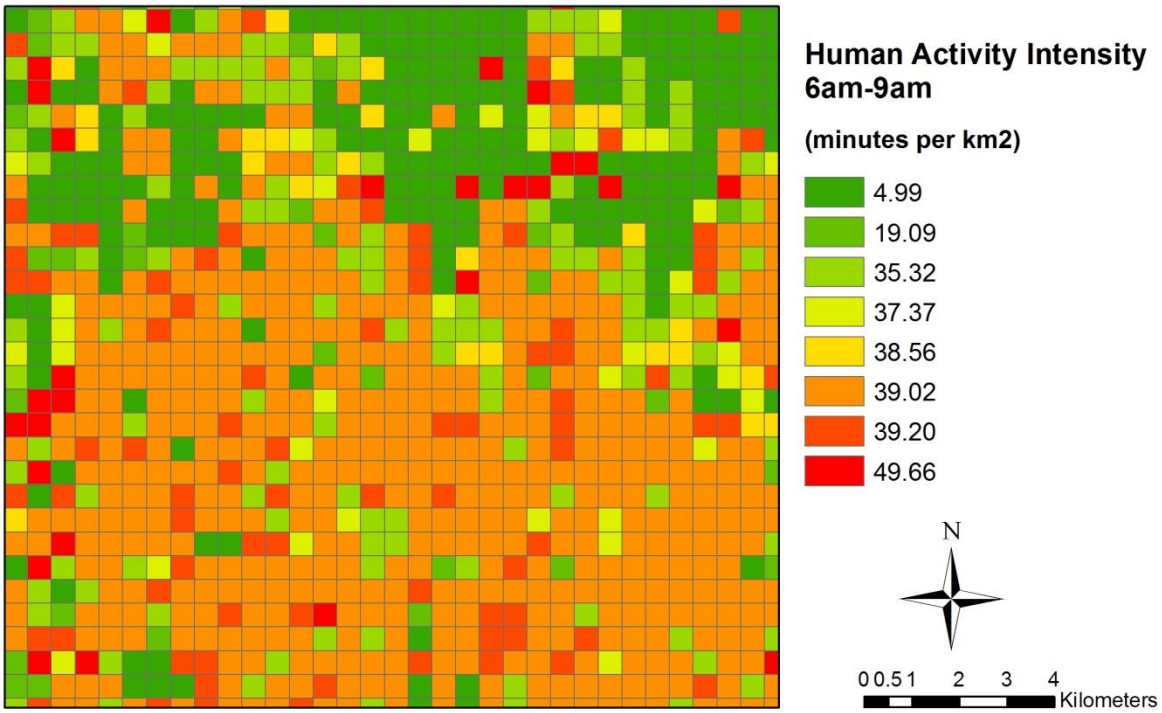

Figure 21. Intensity map of human activity by POI cluster types (one hundred individuals over five weekdays)

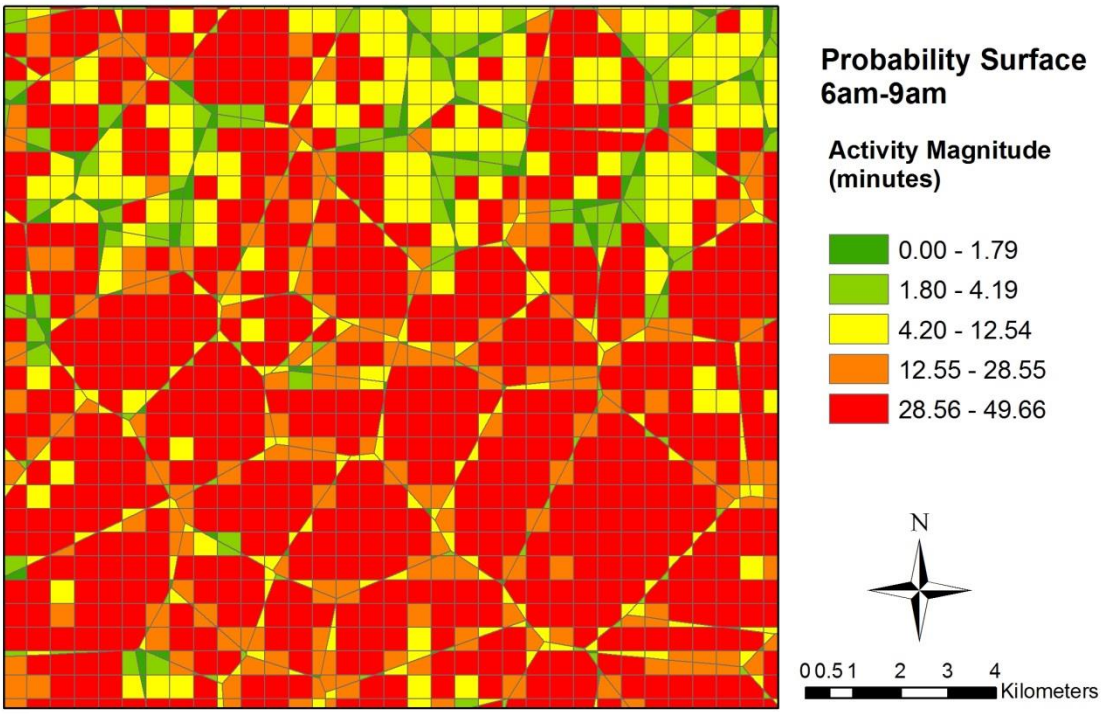

Figure 22. Sum activity magnitude by cellular service network

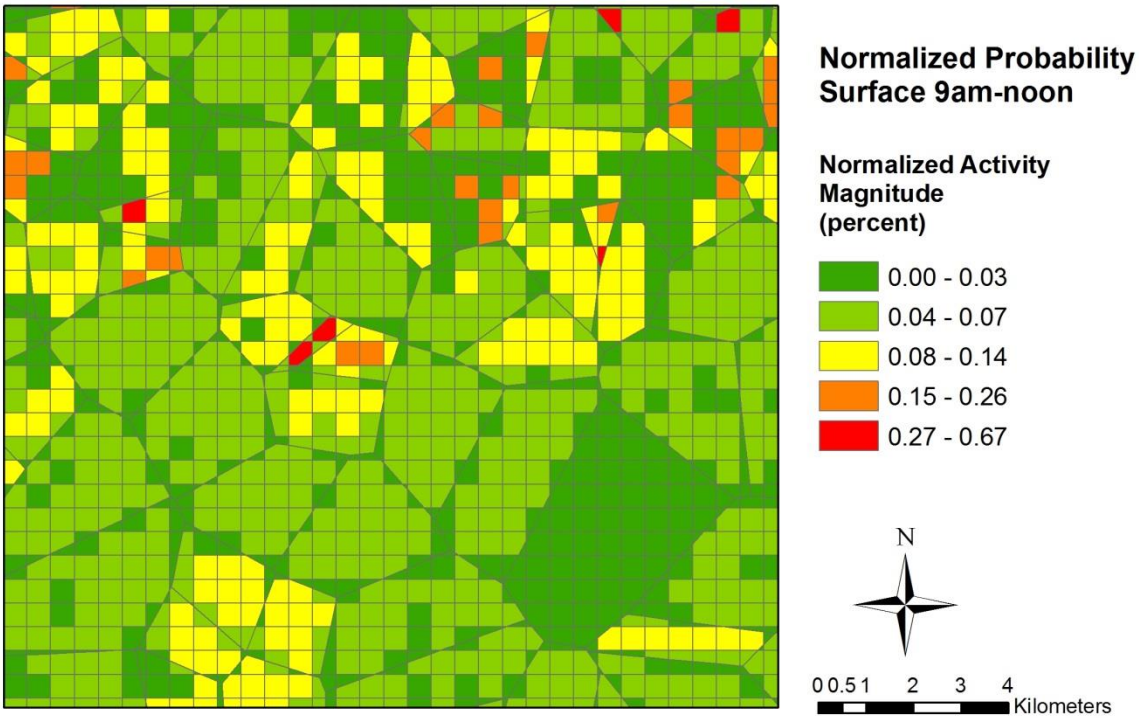

Figure 23. Normalized probability surface by land use

Chapter 5 Discussion

5.1 Land-use method vs. POI method

Two methods were applied with the help of human mobility pattern analysis to improve the spatial accuracy of mobile positioning data, using land use classification and POIs, respectively. When comparing the normalized probability surface produced from the *Land-use method* (Figure 14) and that from the *POI method* (Figure 23), we can see consistent patterns. Figure 14 provides more details within each mobile phone service cell than Figure 23 does because the resulting map from *Land-use method* inherits the fine resolution of 30 by 30 meters from the Landsat ETM+ imagery, while the resolution of the normalized probability surface from *POI method* was resulted from the grids cell size of 500 by 500 meters. This cell size was selected because it has been proved to be successful in analyzing human activity pattern in previous work (Phithakkitnukoon et al. 2010). However, the map results could be sensitive to cell size in my work that has yet to be examined.

Another important difference between these two methods is the form of data they take in every process: the *Land-use method* uses raster calculations while the *POI method* deals with data in vector form. In this context, the *POI method* has advantages over the *Land-use method* on considering boundary conditions and on implementation of locating the estimated CDRs. Another drawback of the *Land-use method* is the robustness of the supervised classification. Although I combined training subsets from different people in this work to avoid biased classification results, there is still concern on how generic or applicable this method is while moving onto other study areas.

5.2 Limitation and future work

A great limitation of this study is the lack of mobile phone positioning data in the study area to evaluate the proposed methods. I have attempted to simulate the mobile phone positionings through power-law distribution as the internal call activity pattern (Barabasi 2005), and used the simulated data to compare the resulting trajectory to the original GPS-based trajectory through methods such as the positional linear feature buffer method developed by Goodchild and Hunter (1997). However, the evaluation results are not very robust. Therefore, real mobile phone positioning data from the study area are still in need to evaluate the two proposed mapping methods.

There is another issue meriting some special attention, especially under the *Land-use method*. It is common sense that GPS signal could be blocked by high buildings in urban areas, which likely leads to an overestimated proportion of activity magnitude at open spaces such as roads and vegetation in this study. This would impact the *POI method* less because it groups POIs into square cells regardless of infrastructure information, rather than having land use as independent categories.

In the *POI method*, other activities POIs still hold a large proportion of all POIs. This would be addressed in greater details in the future work to extract more useful information. Since this work only deals with weekday human activity patterns, the weekends data would also be included into account in the future work.

Reference

Adams, P.M., G.W.B. Ashwell, and R. Baxter. 2003. Location-based services: An overview of standards. *BT Technology Journal* 21, no.1:34-43.

Ahas, Rein, and Ülar Mark. 2005. Location based services: New challenges for planning and public administration? Futures 37, no.6:547-61.

Ahas, Rein, Anto Aasa, Siiri Silm, and Margus. Tiru. 2007a. Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia. In *Information and communication technologies in tourism*, ed. Marianna Sigala, Luisa Mich, and Jamie Murphy. 119-28. Vienna: Springer.

Ahas, Rein, Jaak Laineste, Anto Aasa, and Ülar Mark. 2007b. The spatial accuracy of mobile positioning: Some experiences with geographical studies in Estonia. In *Location Based Services and TeleCartography*, ed. Georg Gartner, William Cartwright, Michael P. *Peterson, 445-60. Springer Berlin Heidelberg.*

Ahas, Rein, Anto Asa, Antti Roose, and Ülar Mark. 2008. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management* 29 no. 3: 469-86.

Ahonen, S, and P. Eskelinen. 2003. Mobile terminal location for UMTS. *IEEE Aerospace and electronic systems magazine* 18, no.2:23-7.

Azevedo, Tiago, Rafael Bezerra, Carlos Campos, and Luis de Moraes. 2009. An analysis of human mobility using real traces. In Wireless Communications and Networking *Conference*. 1-6.

Barabási, Albert-Laszlo. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435: 207-11.

Candia, Julian, Marta C. Gonzalez, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-Laszlo Barabasi. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematica and Theoretical* 41, no. 22:1-16.

Clauset, Aaron, and Nathan Eagle. 2007. Persistence and periodicity in a dynamic proximity network. In *Proceedings of Discrete Mathematics and Theoretical Computer Science Workshop on Computational Methods for Dynamic Interation Networks*.

Eagle, Nathan, Alex Pentland, and David Lazer. 2007. Inferring friendship network structure using mobile phone data. PNAS 106, no.36:15274-8

Eicher, Cory L., and Cythia A. Brewer. 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science* 28, no.2:125-38

González Marta C., Cesar A. Hidalgo, and Albert-Laszlo Barabási. 2008. Understanding individual human mobility patterns. Nature 453:779-82.

Hunter, Gary J., Michael F. Goodchild. 1997. Modeling the Uncertainty of Slope and Aspect Estimates Derived from Spatial Databases. Geographical Analysis 29, no. 1:35-49.

He, Chun-yang, Pei-jun Shi, Jin Chen, and Yu-yu Zhou. 2001. A study on land use/cover change in Beijing area. *Geographical Research* 20, no.6: 679-88.

Janelle, Donald G. 1995. Metropolitan expansion, telecommuting and transportation. In The *geography of urban transportation*, ed. Susan Hanson, 402-34. New York: Guilford Press.

Kang, Chaogui, Yu Liu, Xiujun Ma, Lun Wu. 2012. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology* 19: 3-21.

Kwan, Mei-po. 2007. Mobile communications, social networks, and urban travel: hypertext as a new metaphor for conceptualizing spatial interaction. *The Professional Geographer* 59 no.4:434–446.

Li, S., J. Wang, Y. Bi, Y. Chen, M. Zhu, S. Yang, and J. Zhu. 2005. A review of methods for classification of remote sensing images. *Remote Sensing for Land & Resources* 64, no.2: 1-5

Lu, Yongmei, Yu Liu. 2012. Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems* 36 no.2:105-8.

Minch R. P. 2004. Privacy issues in location-aware mobile devices. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*.

Mokhtarian, Patricia L, Ravikumar Meenakshisundaram. 1999. Beyond tele-substitution: disaggregate longitudinal structural equations modeling of communication impacts. *Transportation Research Part C: Emerging Technologies* 7: 33-52.

Onnela J.-P, J. Saramaki, J. Hyvonen, G. Szabó, D. Lazer, K. Kaski, J. Kertesz, and A.-L Barabasi. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104: 7332-6.

Paola, J. and Schowengerdt, R. 1995. A detialed comparison of backpropagation Neural Network and Maximum-Likelihood Classifiers for urban land use classification. *IEEE Transactions on Geoscience and Remote Sensing* 33, no.4: 981-996.

Phithakkitnukoon, Santi, Teerayut Horanont, Giusy D. Lorenzo, Ryosuke Shibasaki, Carlo Ratti. 2008. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In Human behavior understanding, ed. Albert Salah et al. 14-25. Springer Berlin Heidelberg.

Ratti, Carlo, Riccardo M. Pulselli, Sarah Williams, and Dennis Frenchman. 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33: 727–48.

Schulz, Daniel, Sebastian Bothe, and Christine Körner. 2012. Human mobility from GSM data - A valid alternative to GPS? In *The Nokia Mobile Data Challenge Workshop*.

Silm, Siiri, and Rein Ahas. 2010. The seasonal variability of population in Estonian municipalities. *Environment and Planning A* 42, no.10:2527-46.

Song, Chaoming, Zehui Qu, Nicholoas Blumm, and Albert-Laszlo Barabási. 2010. Limits of predictability in human mobility. Science 327, no. 5968:1018-21.

Wang, Pu., Marta C. González, Cesar A. Hidalgo, and Albert-Laszlo Barabási. 2009. Understanding the spreading patterns of mobile phone viruses. *Science* 324:1071.

Wang, Pu, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, and Marta C. Gonzalez. 2012. Understanding road usage patterns in urban areas. *Scientific Report* 2.

Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. 2012. Quantifying the impact of human mobility on malaria. *Science* 338 no.6104: 267-70.